Extraction de données
Approche du projet SoDUCo

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

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

Social Dynamics in Urban Context
PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

(Ré)utiliser cette présentation

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…

Non-Commerçans. (Paris). 269

Chardin, R. Pavée, 16. — B. C.
Chardin, R. Michel Lepelletier, 21.
Chardon, (Ve.) R. S. Marc, 15.
Chevíron, R. Chapon, 13.
Chimay, (Mme.) R. de Varennes, 31.
Choart-Duplessis, R. de Turenne, 31.

AMADOU ET ALLUMETTES. — Pour les ALLUMETTES OXIGÉNÉES. Voyez BRIQUETS PHYSIQUES.

DARRAS (Thomas), r. de la Vieille-Monnaie, 10. Briquets et veilleuses, mèches à quinquets, mèches soufrées, pierres, agaric de chêne, liége en planches, bouchons.

GALLIENNE jne., r. de la Heaumerie, 3. Brûle-tout, boites à briquet, mèches à vin et à quinquet, veilleuses mèches, soufrées; pierres, agaric, bouchons, liége.

LEROY, r. Aubry-le-Boucher, 43.

BAUDOYER (place).

IX Arr. *Hôtel-de-Ville.*) ← Rue Tixeranderie, pourtour St-Gervais, Saint-Antoine et Renaud-Lefèvre.

1 Lissoty (Vve), *vins.*
2* Privé, *distillateur.*
Lemoine-Cluzel et Leroy, *nouveautés.*
Chantrier, *courtier-gourm.*
26* Longpré aîné, *bijoutier en or et argent.*
Saint-Omer, *émailleur.*
Cellier (A.), *graveur-ciseleur.**
Rousseau (J.), *bijoutier en or.**
Benoît, *orfèvre-fabr.*
Lérésy, *doreur.*
30 Bouton, *fab. de cuir vernis.**
31 Pardon, *vins.*
Bourguille, *fabr. de presses.*
Vaudain, *passementier.*
Fininoinc, *bronze doré.*
Rabé aîné, *fabr. de ballons.**
Gaulin, *chapelier.*
Moisy, *tabletier.*
49* Cendrier aîné, *prop.*
Desmarests, *fab. boîtes d'emballage.*
Ferrand, *lapidaire.*
7 Ecole communale de jeunes filles.
Berthelot, *vins.*
6 Verstaen, *serrurier-mécanicien.*
8 Michel, *brossier.*
9 Labottiere, *serrurier.*
10 Sacrez, *vins.*
12 Baudoin, *épic.*
13 Lejard, *clouteries et crépins.*
14 Baduel (Vve), *fab. de et tapisseries.*
10 Lainé jeune, *vins.*
Jumelles omnibus et entreprise générale des Omnibus.
11 Melouzay, *vins en gros,* et à Bercy, Port, 31.
12 Combaud, *coiffeur.*
Monmain (P.), *vins en gros.*
13 Dufailly, *sculpt. fabr. de carton-pierre.*

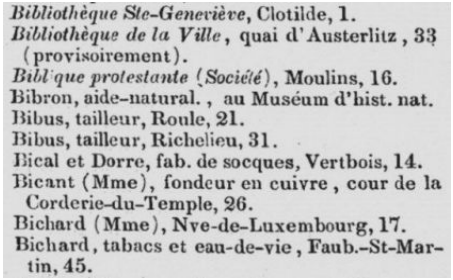






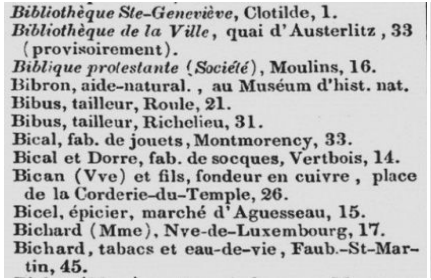


Un instantané des acteurs commerciaux à un instant



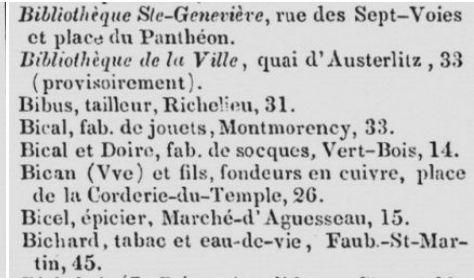
Bibliothèque Ste-Geneviève, Clotilde, 1.
Bibliothèque de la Ville, quai d'Austerlitz, 33 (provisoirement).
Biblique protestante (Société), Moulins, 16.
Bibron, aide-natural., au Muséum d'hist. nat.
Bibus, tailleur, Roule, 21.
Bibus, tailleur, Richelieu, 31.
Bical et Dorre, fab. de socques, Vertbois, 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-Martin, 45.

Didot 1841a - page 95



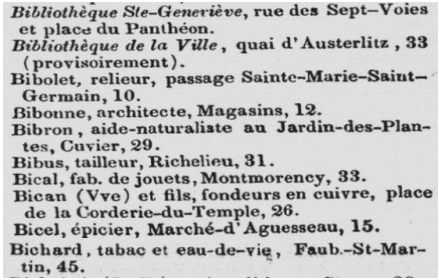
Bibliothèque Ste-Geneviève, Clotilde, 1.
Bibliothèque de la Ville, quai d'Austerlitz, 33 (provisoirement).
Biblique protestante (Société), Moulins, 16.
Bibron, aide-natural., au Muséum d'hist. nat.
Bibus, tailleur, Roule, 21.
Bibus, tailleur, Richelieu, 31.
Bical, fab. de jouets, Montmorency, 33.
Bical 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-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

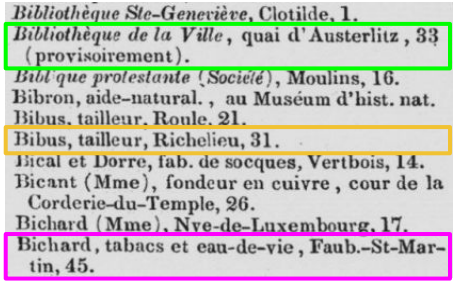


Bibliothèque Ste-Geneviève, rue des Sept-Voies et place du Panthéon.
Bibliothèque de la Ville, quai d'Austerlitz, 33 (provisoirement).
Bibolet, relieur, passage Sainte-Marie-Saint-Germain, 10.
Bibonne, architecte, Magasins, 12.
Bibron, aide-naturaliste au Jardin-des-Plantes, Cuvier, 29.
Bibus, tailleur, 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.

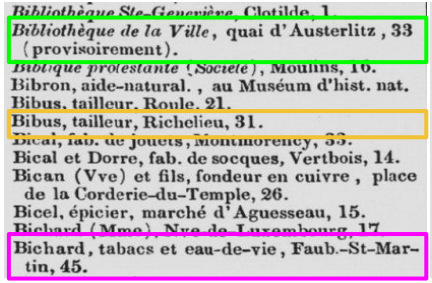
Didot 1844a - pages 125-126

Un instantané des acteurs commerciaux à un instant

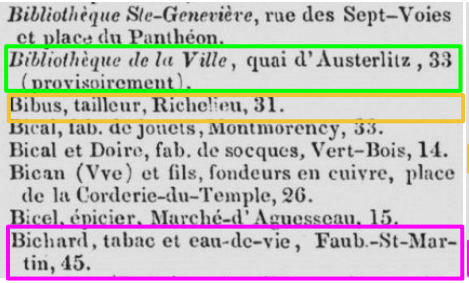
Grande **redondance** d'information intra- et inter-annuaires.


Bibliothèque Ste-Geneviève, Clotilde, 1.
Bibliothèque de la Ville, quai d'Austerlitz, 33 (provisoirement).
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-Martin, 45.

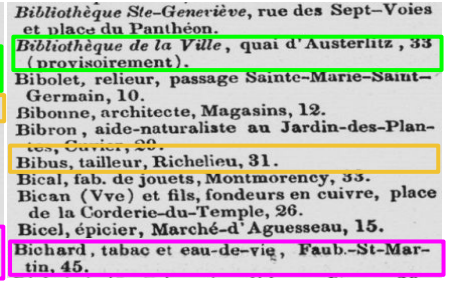
Didot 1841a - page 95


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-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.

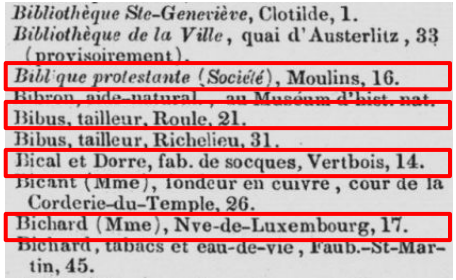
Didot 1843a - page 129


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, 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.

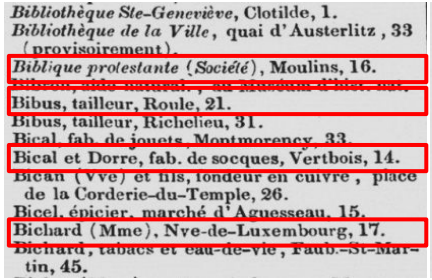
Didot 1844a - pages 125-126

Un instantané des acteurs commerciaux à un instant

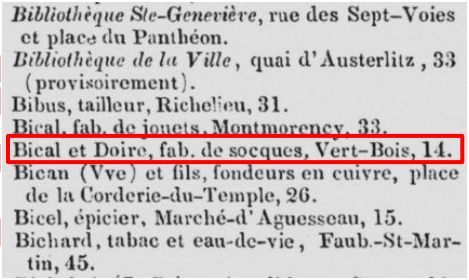
Des entrées **disparaissent** au cours du temps.


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

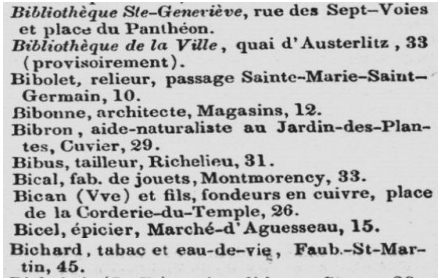
Didot 1841a - page 95


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

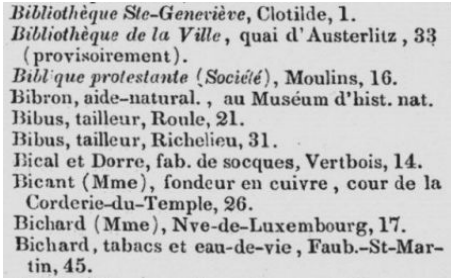
Didot 1843a - page 129

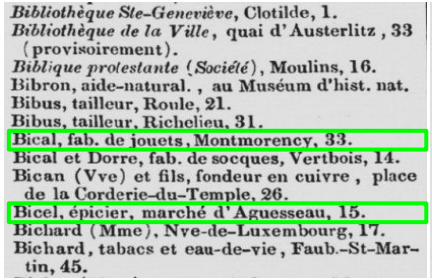

Bibliothèque Ste-Geneviève, rue des Sept-Voies et place du Panthéon.
Bibliothèque de la Ville, quai d'Austerlitz, 33 (provisoirement).
Bibolet, relieur, passage Sainte-Marie-Saint-Germain, 10.
Bibonne, architecte, Magasins, 12.
Bibron, aide-naturaliste au Jardin-des-Plantes, Cuvier, 29.
Bibus, tailleur, 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.

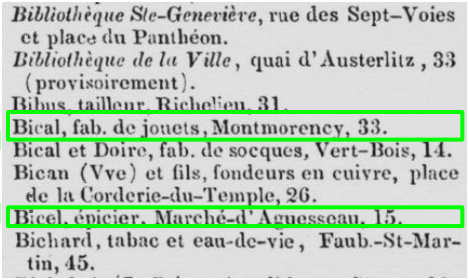
Didot 1844a - pages 125-126

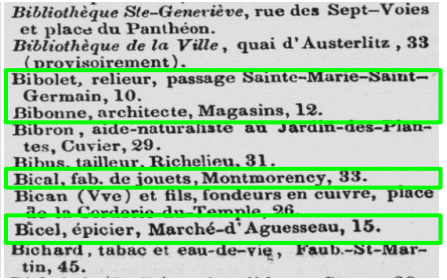
Un instantané des acteurs commerciaux à un instant

D'autres **apparaissent**.


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


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.

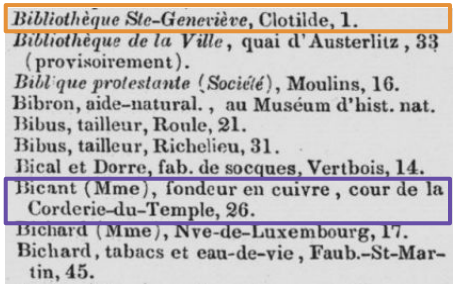

Bibliothèque Ste-Geneviève, rue des Sept-Voies et place du Panthéon.
Bibliothèque de la Ville, quai d'Austerlitz, 33 (provisoirement).
Bibus, tailleur, Richelieu, 31.
Bical, fab. de jouets, Montmorency, 33.
Bical et Doire, fab. de socques, Vert-Bois, 14.
Bican (Vve) et fils, fondeurs en cuivre, place 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, 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.

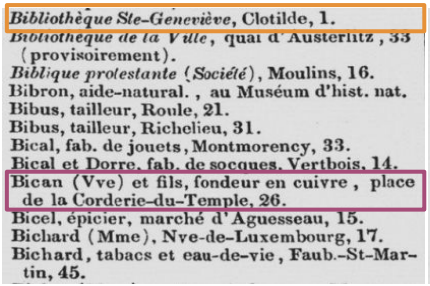
Didot 1841a - page 95 *Didot 1842a - page 117* *Didot 1843a - page 129* *Didot 1844a - pages 125-126*

Un instantané des acteurs commerciaux à un instant

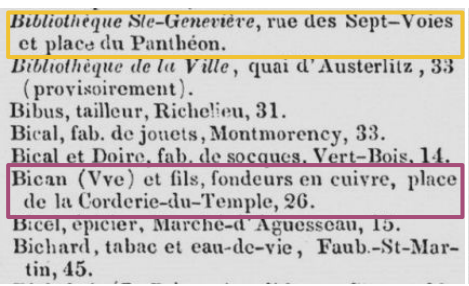
Et certaines **changent**.


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-Martin, 45.

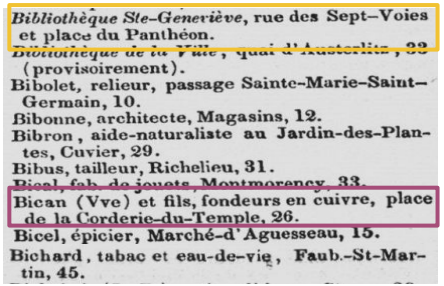
Didot 1841a - page 95


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-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.

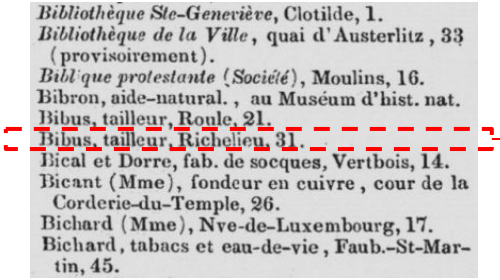
Didot 1843a - page 129


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

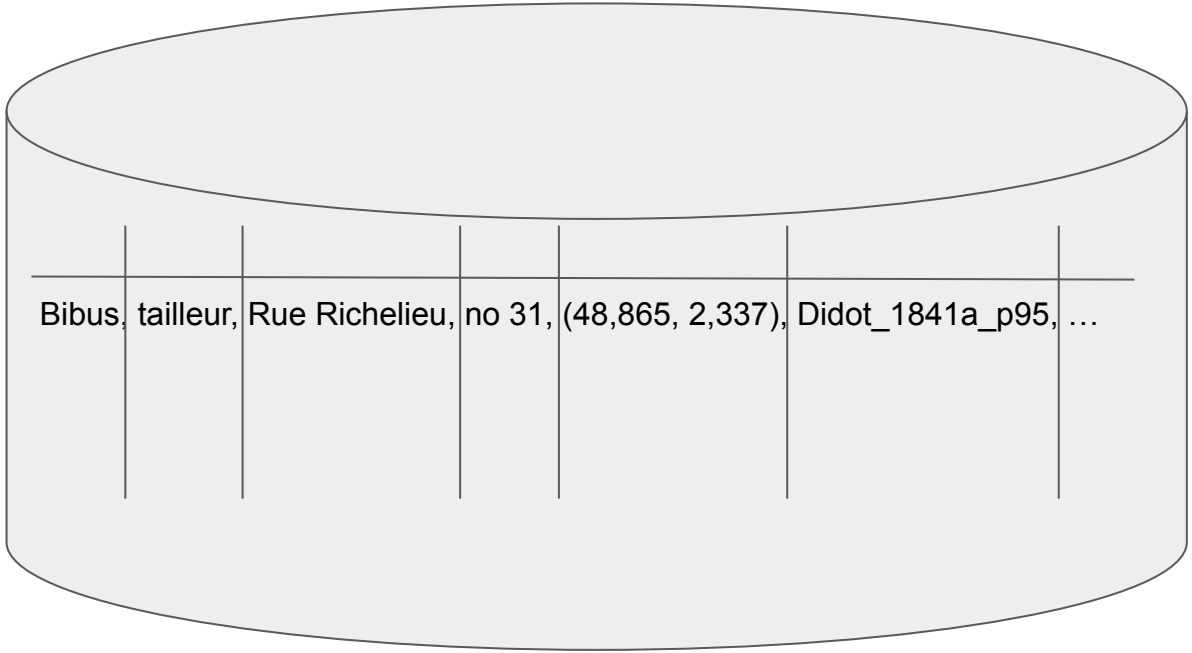
Didot 1844a - pages 125-126

Notre objectif d'extraction de données

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

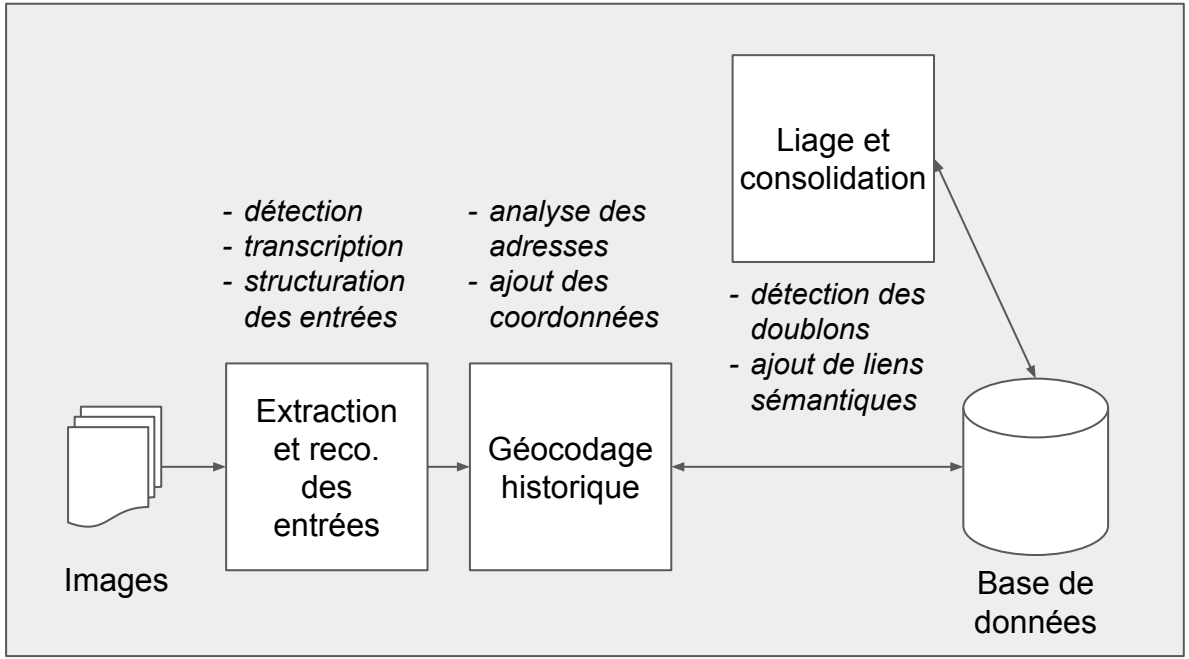

Bibliothèque Ste-Geneviève, Clotilde, 1.
Bibliothèque de la Ville, quai d'Austerlitz, 33 (provisoirement).
Bibl'que protestante (Société), Moulins, 16.
Bibron, aide-natural., au Muséum d'hist. nat.
Bibus, tailleur, Roule, 21.
Bibus, tailleur, Richelieu, 31.
Bical et Dorre, fab. de socques, Vertbois, 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



Vue d'ensemble du processus

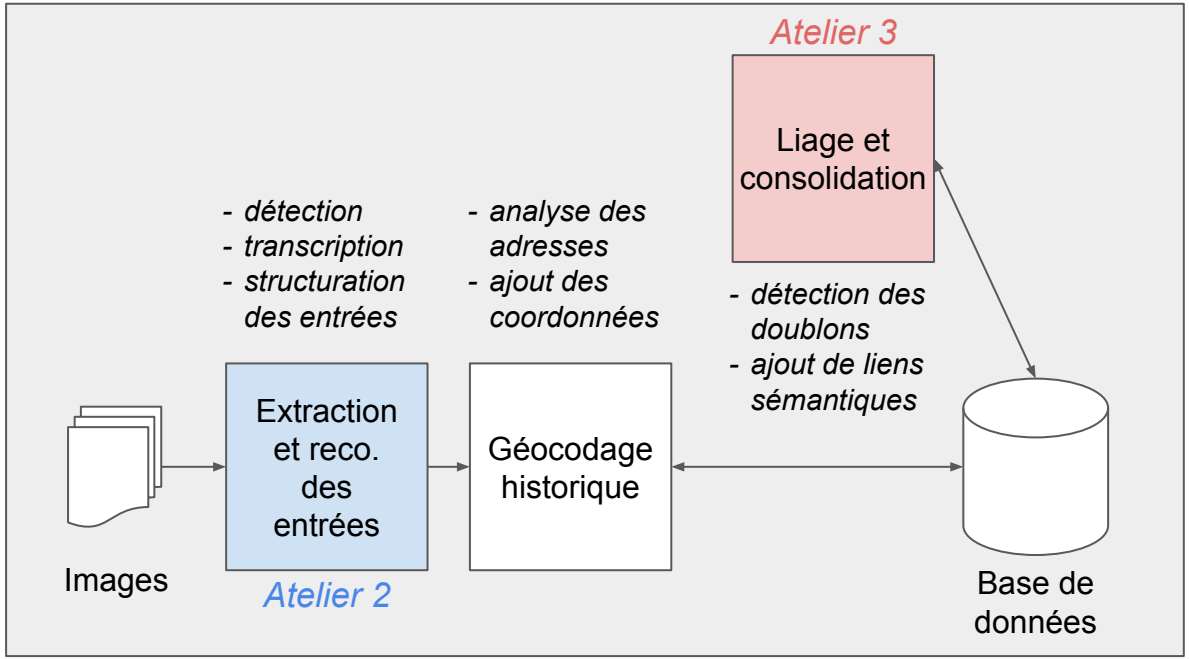
Définition du corpus



Analyse des données

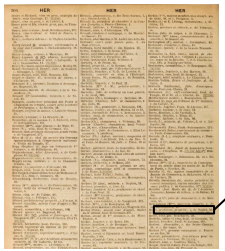
Vue d'ensemble du processus

Définition du corpus



Analyse des données

Extraction et reconnaissance des entrées



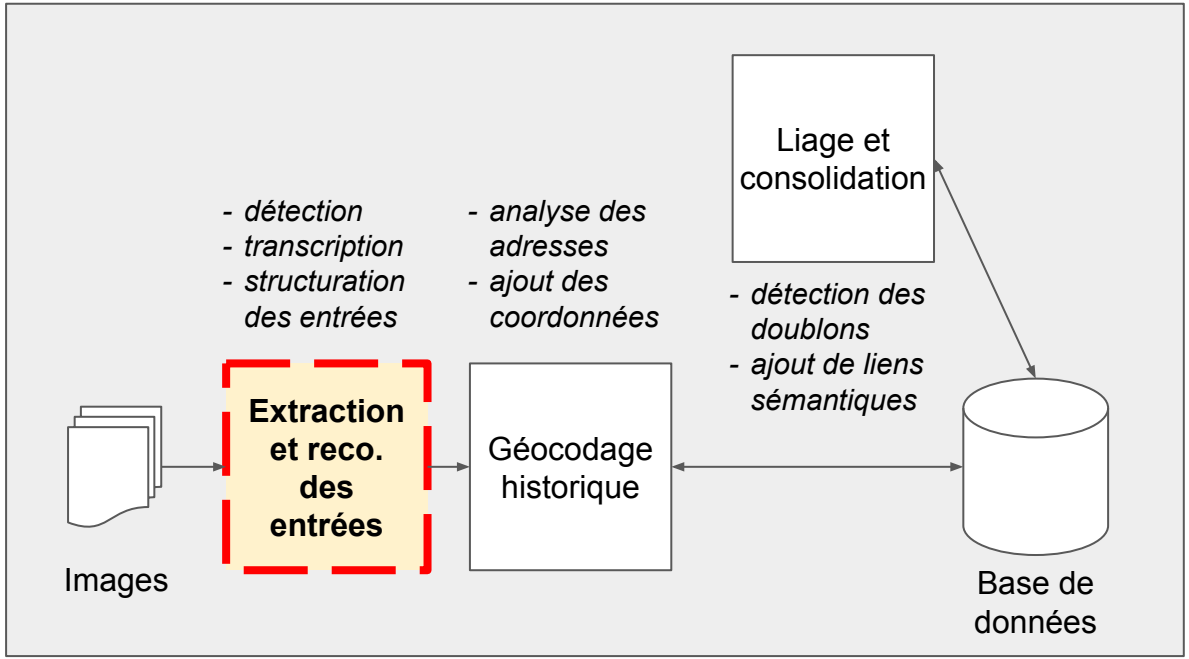
Lemonnyer, *plomberie*, r. de Bondy, 86, et r. Bouchardon, 1.

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

Définition du corpus

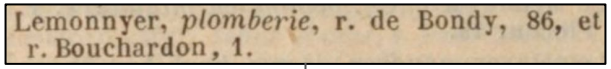


Analyse des données

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



Transcription

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

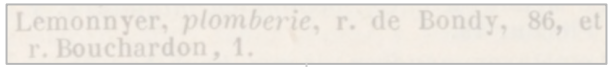
Extraction des
informations clés

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



Transcription

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

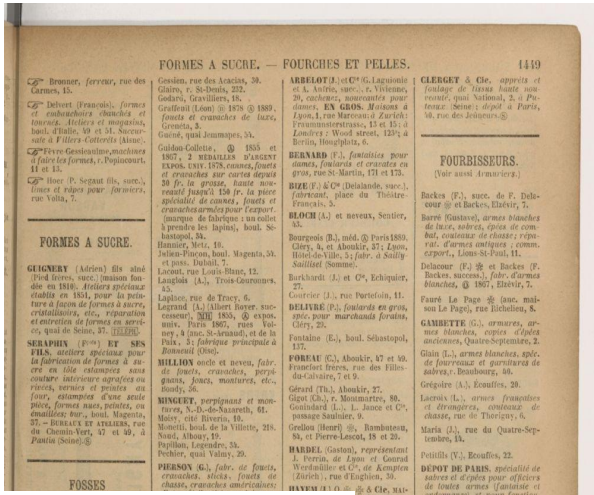
Extraction des
informations clés

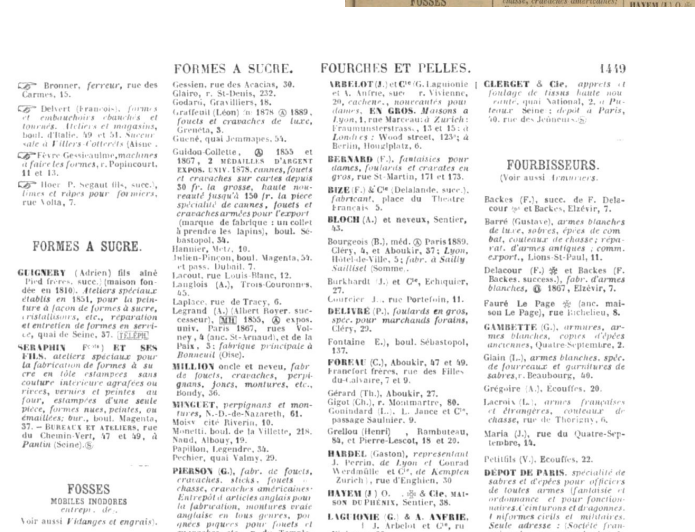
PERSONNE ACTIVITÉ RUE
Lemonnyer, plomberie, r. de Bondy,
NUMÉRO RUE NUMÉRO
86, et r. Bouchardon, 1.

Amélioration de la qualité d'image

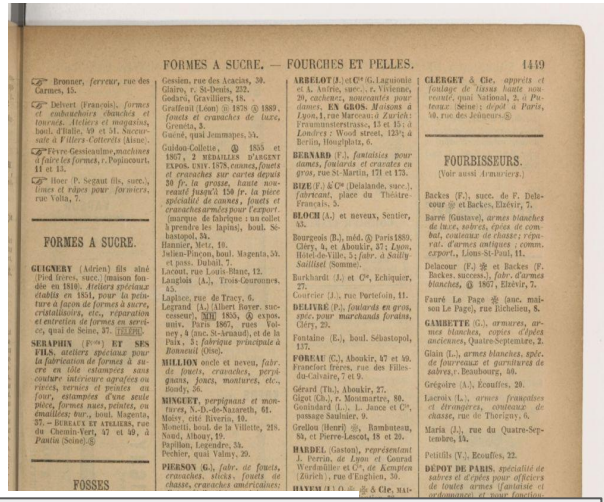
- Réduire la variance des entrées pour la suite des traitements

✓ Amélioration de contraste / suppression de fond
✓ Rotation / Shear





Amélioration de la qualité d'image


FORMES A SUCRE. — FOURCHES ET PELLES. 1449

FORMES A SUCRE.

FOSSES

FOURBISSEURS.

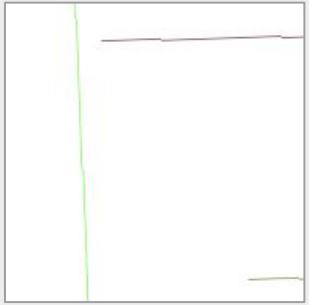
- Réduire la variance des entrées pour la suite des traitements

✓ 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



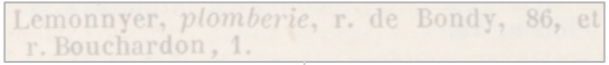
Lignes détectées

Image redressée

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



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

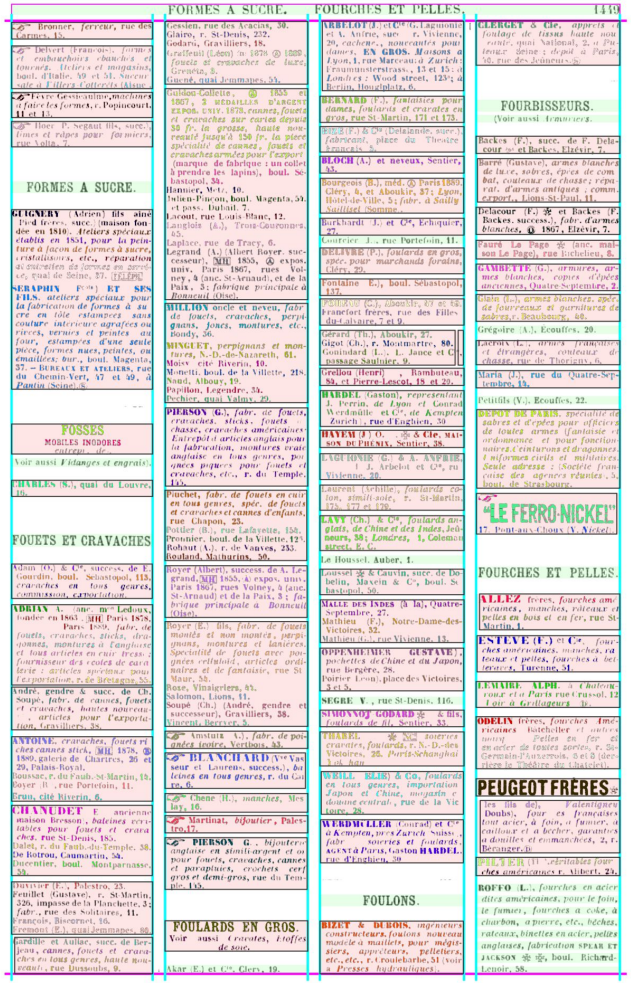
1. Séparation en **blocs** (XY cuts, smearing)
   et Classification des régions (entête, titre…)
2. Détection des **lignes** (watershed)
3. Regroupement des lignes en **entrées** (HMM)

Avantages

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

Limites

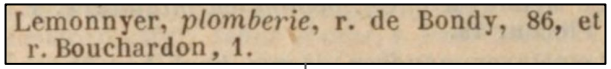
- *Extraction d'entrées intégrée*
- Pas d'exploitation de l'information textuelle
- Pas de gestion du multi-pages
- Sensible au bruit, d'où le nettoyage en amont



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



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.

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

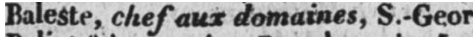
Systèmes testés

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

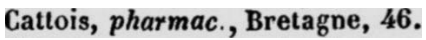
Pas de ré-entraînement.

Bottin 1820

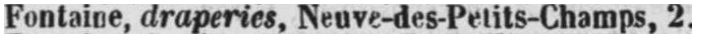
Bottin 1827



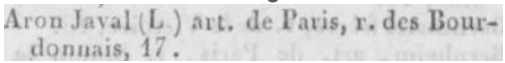
Bottin 1837



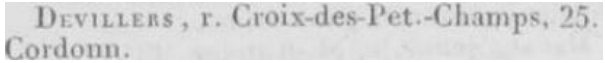
Bottin 1854



Cambon Almgene 1841

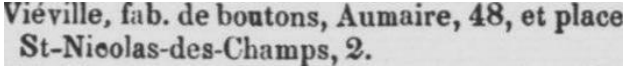


Deflandre 1828



Deflandre 1829

Didot 1851



Reconnaissance du texte (OCR)

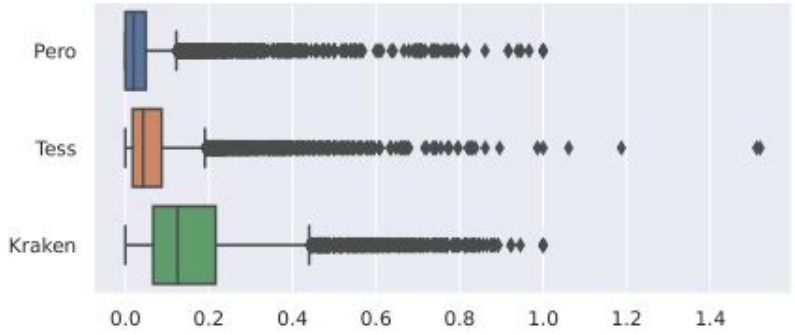
Résultats

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

	PERO OCR	Tesseract	Kraken
CER	3.78%	6.56%	15.72%

Autres systèmes libres compétitifs testés depuis

- Microsoft UniLM/TrOCR
- Mindee docTR

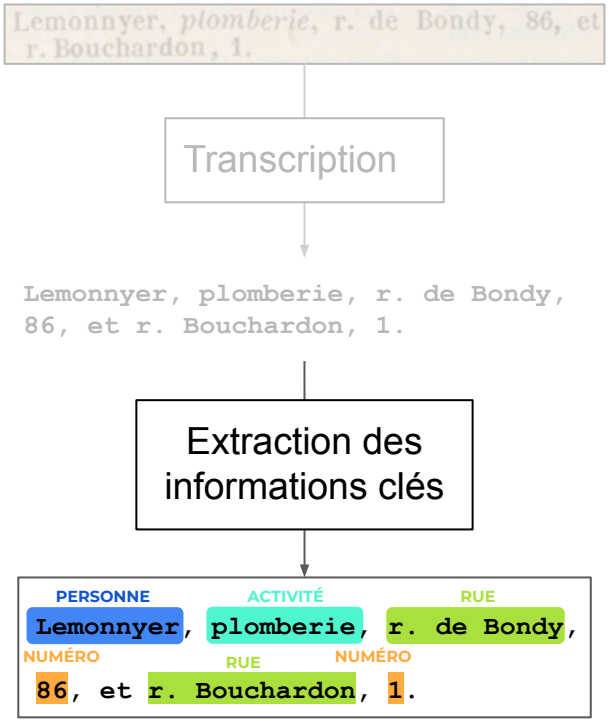


⇒ OCR plutôt performant, résultats exploitables

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



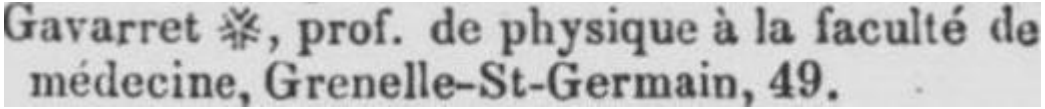
Reco. d'entités nommées (*Named Entity Recognition*)

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.

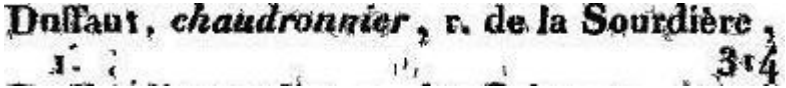
Exemple:

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


Daflant, *chaudronnier*, r. de la Sourdière,

```
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 :

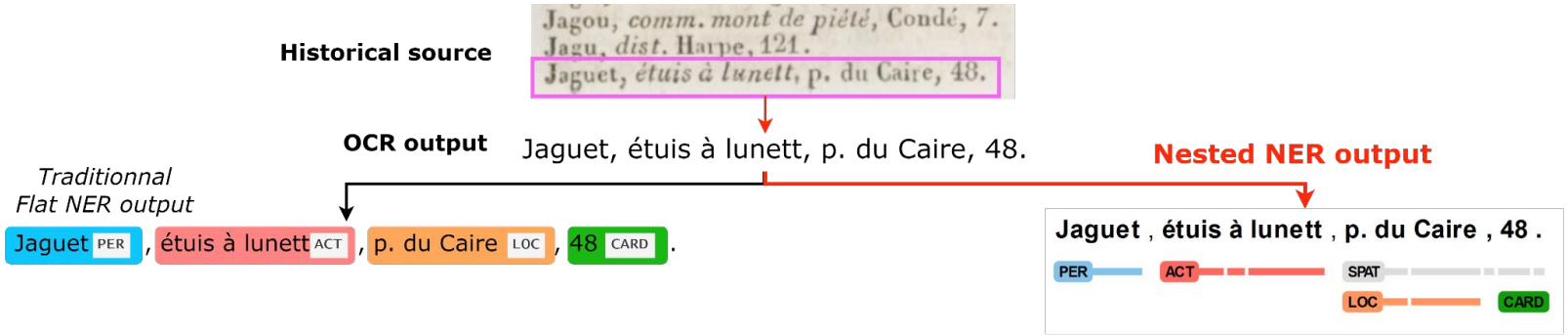
N. Abadie, E. Carlinet, J. Chazalon et B. Duménieu, "A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories", in Proc. DAS 2022, https://github.com/soduco/paper-ner-bench-das22

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 :

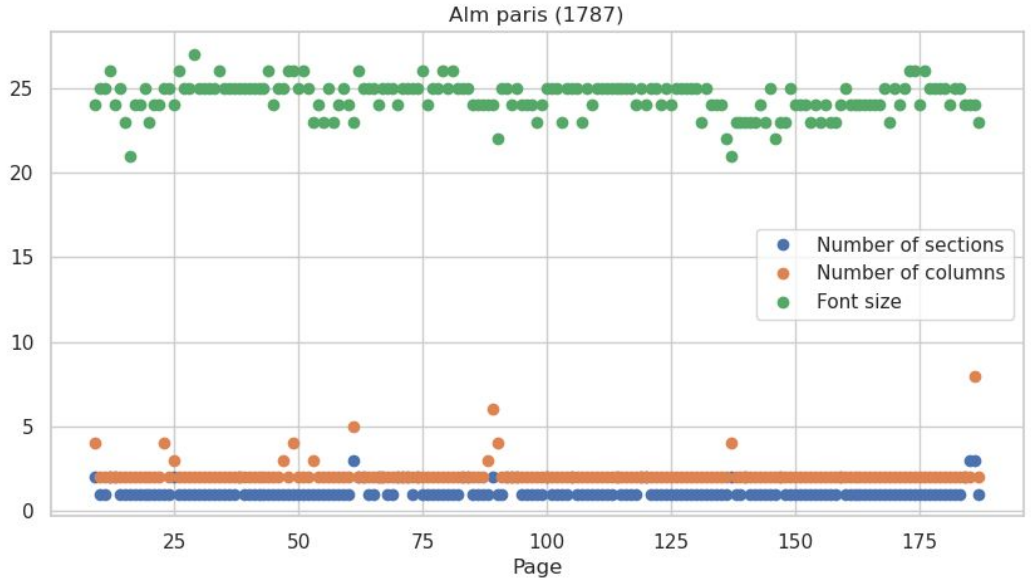
PERSON [ACT] LOC [CARDINAL]

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 majoritairements textuels, ordre de lecture simple

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

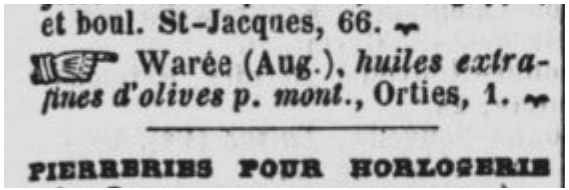

et boul. St-Jacques, 66.
Warée (Aug.), *huiles extra-fines d'olives p. mont.*, Orties, 1.
PIERRERIES POUR HORLOGERIE

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 attirante.

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.

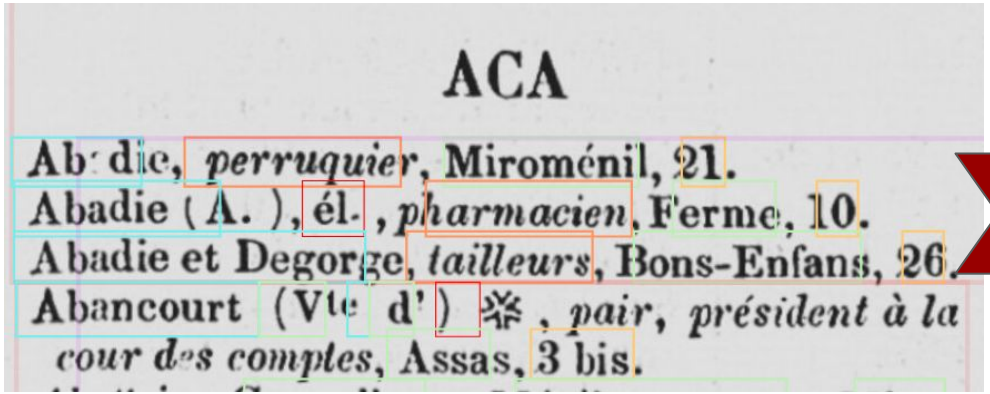
🏁 Au final…

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

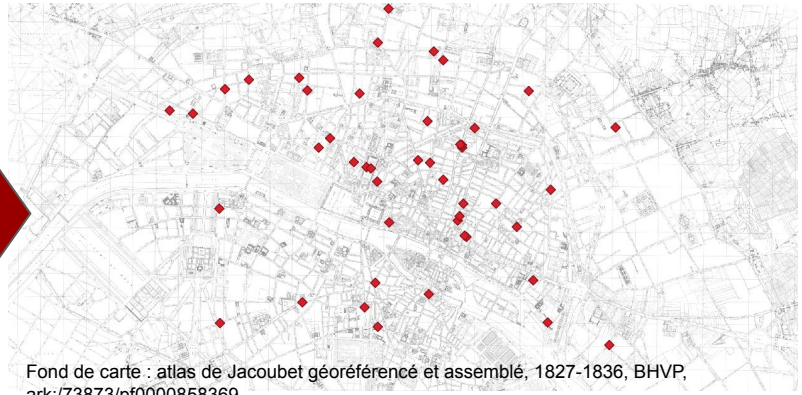
(Lamy_1840) Annuaire général du commerce, judiciaire et administratif de France et des principales villes du monde, v. 128.



ACA

Ab: die, *perruquier*, Miroménil, 21.
Abadie (A.), él., *pharmacien*, Ferme, 10.
Abadie et Degorge, *tailleurs*, Bons-Enfans, 26.
Abancourt (Vte d') ✲, *pair*, *président à la cour des comptes*, Assas, 3 bis.

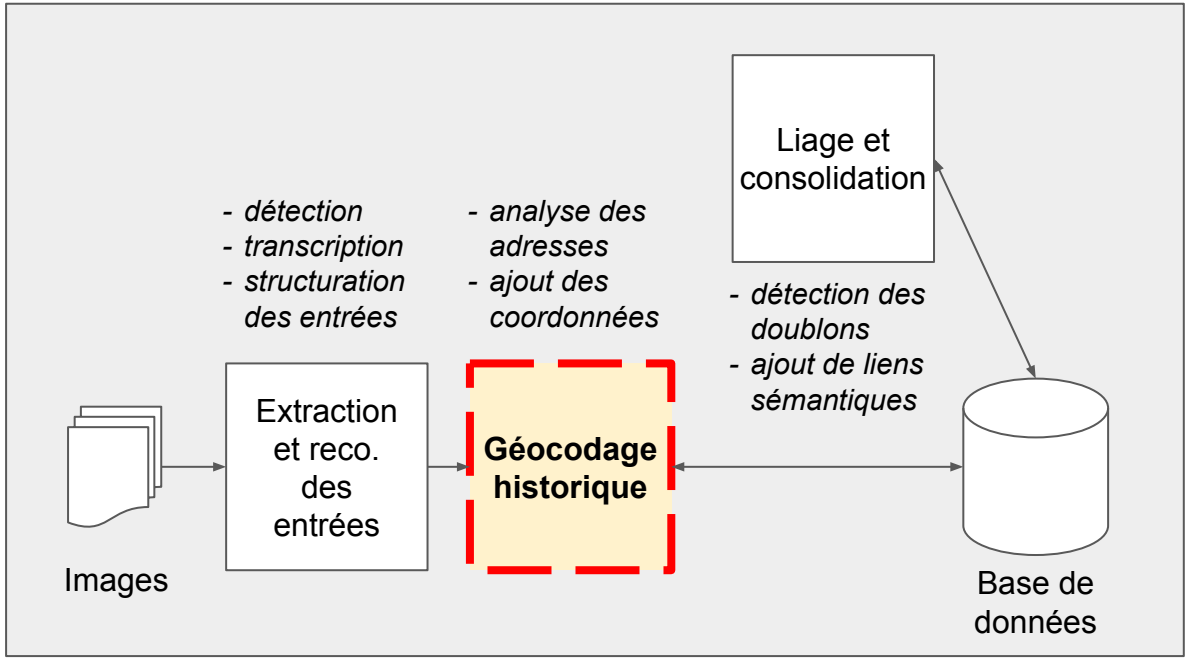
https://api.geohistoricaldata.org/directories/export.geojson?"source.pdf_id"=eq.Lamy_1840&"source.pdf_view"=eq.0128&



Fond de carte : atlas de Jacoubet géoréférencé et assemblé, 1827-1836, BHVP, ark:/73873/pf0000858369.

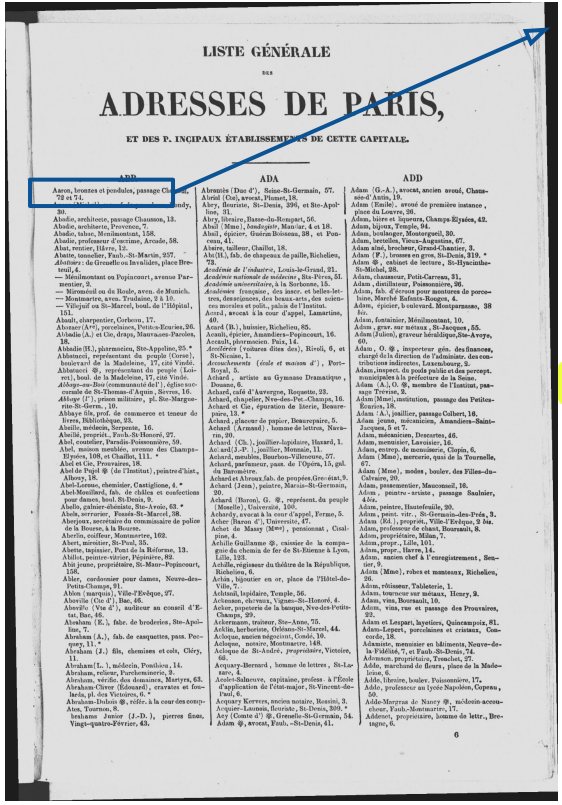
Vue d'ensemble du processus

Définition du corpus



Analyse des données

Chaîne de traitement


LISTE GÉNÉRALE

ADRESSES DE PARIS,

ET DES P. INCIPAUX ÉTABLISSEMENS DE CETTE CAPITALE.

ADA

ADA

ADD

#1 OCR

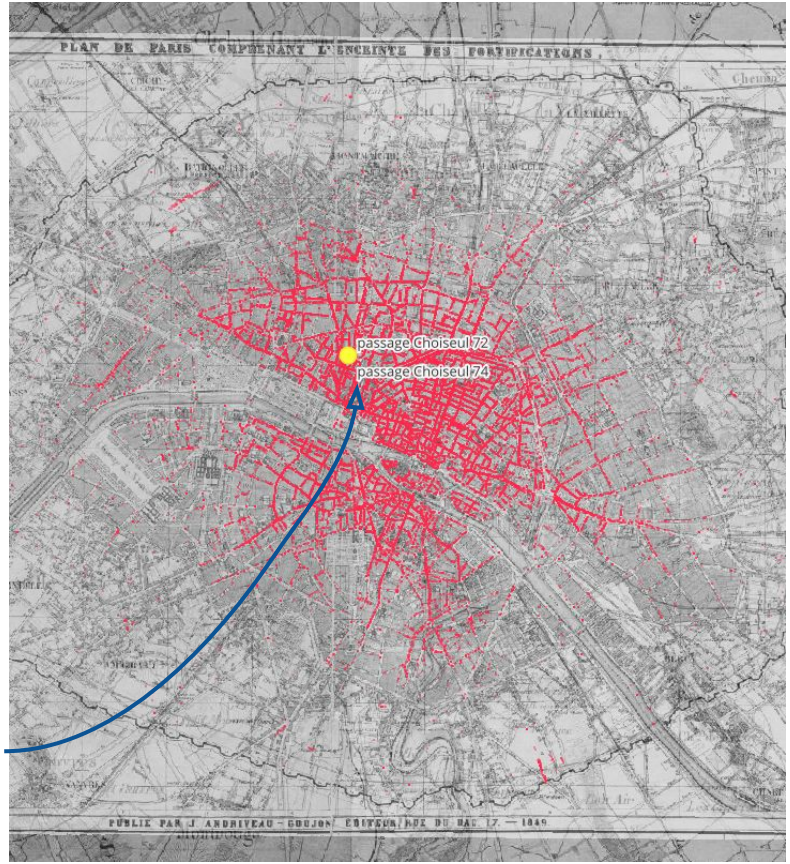
```
> Aaron, bronzes et pendules,
passage Choiseul, 72 et 74.
```

#2 NER

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

#3 Restructuration

```
Address #1 : passage Choiseul 72
Address #2 : passage Choiseul 74
```



(Didot_1851) Annuaire général du commerce, de l'industrie, de la magistrature et de l'administration 185. Liste alphabétique, 341 pages, 700 042 entrées reconnues.

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

adresse
"passage Choiseul, 72"

année visée (scalaire ou intervalle)
[1850,1852]

Recherche multicritères dans les données vectorielles des plans :

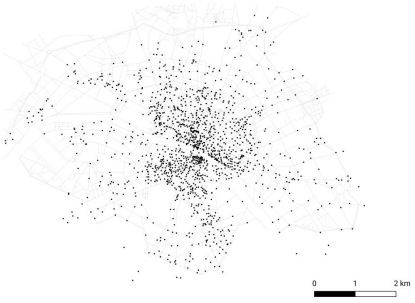
- matching flou du nom + numéro de rue
- distance temporelle

Parsing LibPostal
"passage Choiseul" "72"

elasticsearch

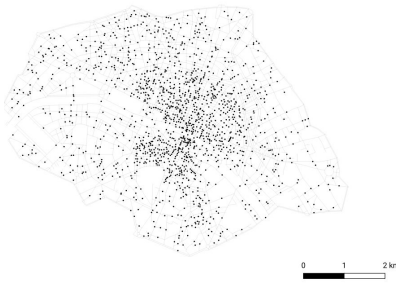
Pelias
https://pelias.io/

Adaptation géohistorique du géocodeur libre et *open source* PELIAS



Atlas de Verniquet, 1784-1795
Filaire des rues

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

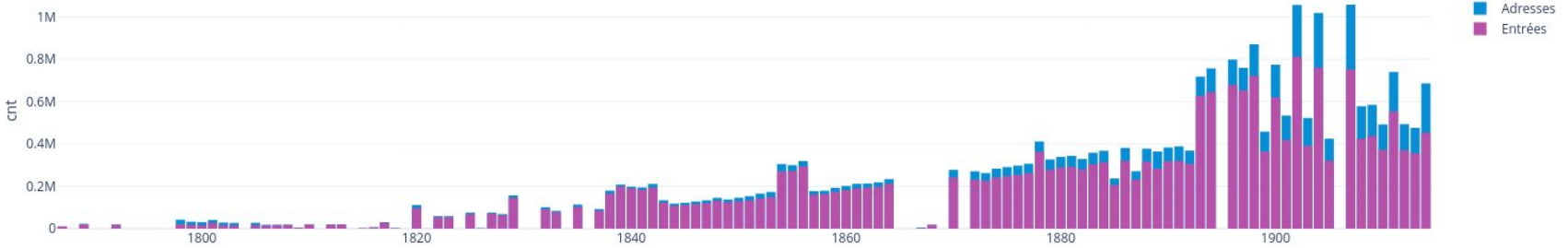


Plan Andriveau-Goujon, 1849
Filaire des rues

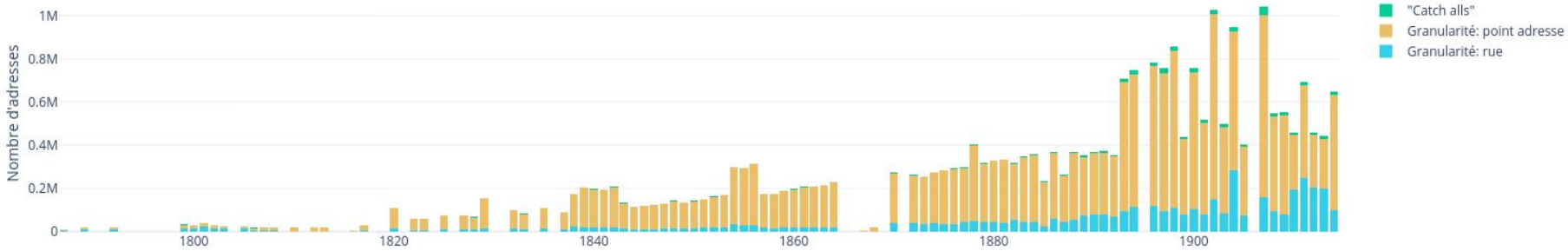
Atlas municipal des 20 arrds., 1888
Filaire des rues & points adresse

Géocodage des annuaires : aperçu des résultats

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



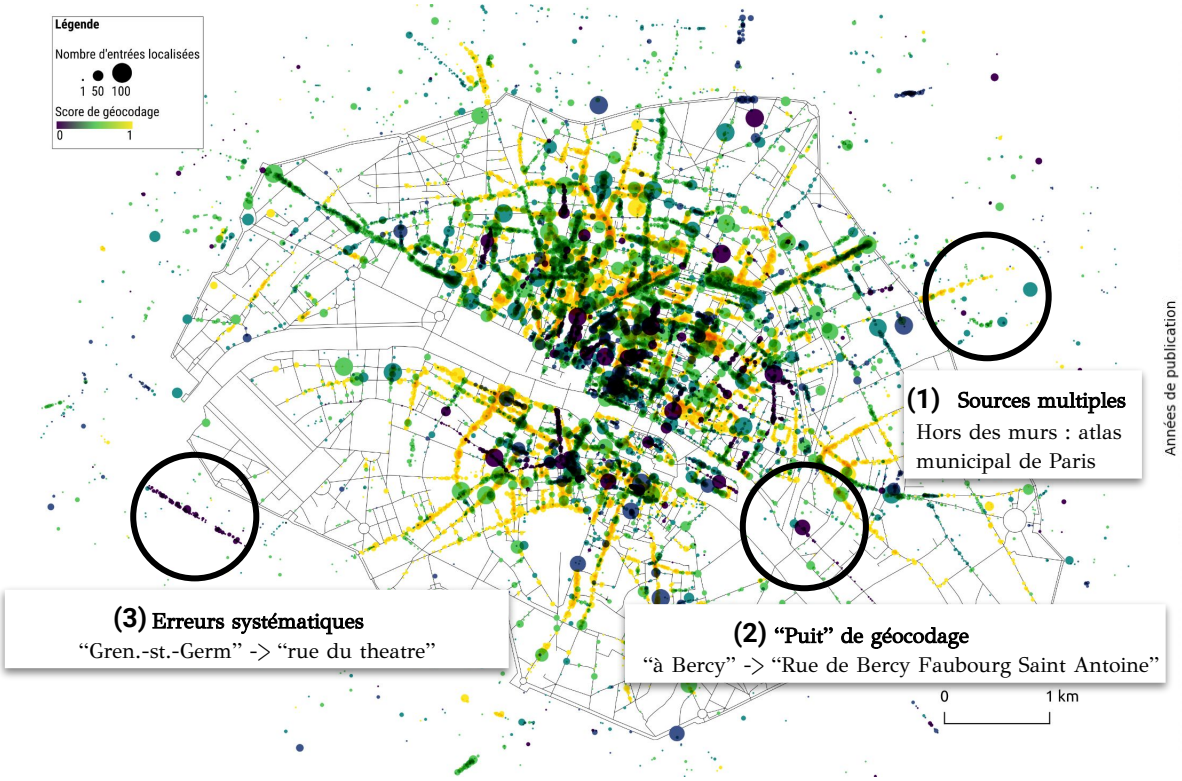
Nombre d'entrées & d'adresses extraites, par année de publication des annuaires.



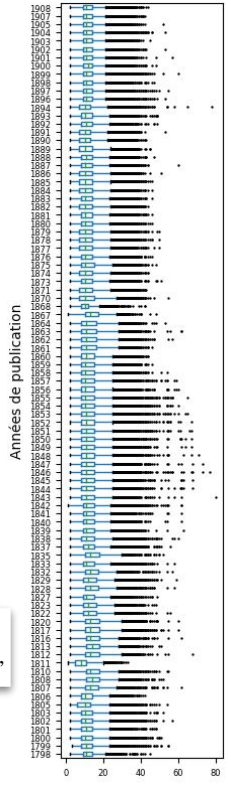
Adresses extraites par années de publications de annuaires et nombre d'adresses géocodées.

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



Nombre de caractères

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

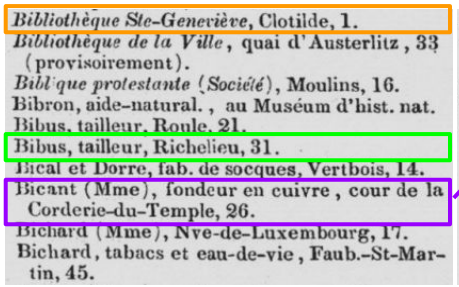
avenue de Neuilly, sur la pelouse près la rue du chemin de Versailles

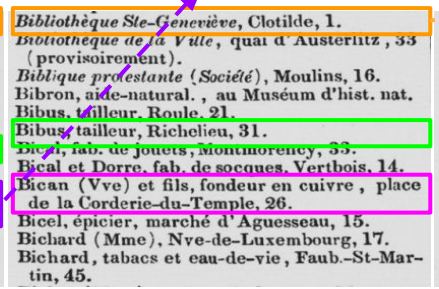
Liage et consolidation

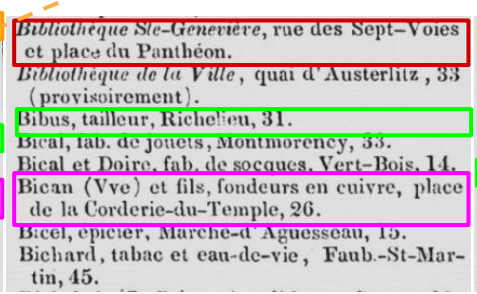
même personne

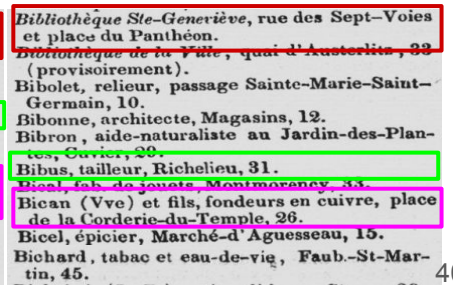
succession

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, 23.
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

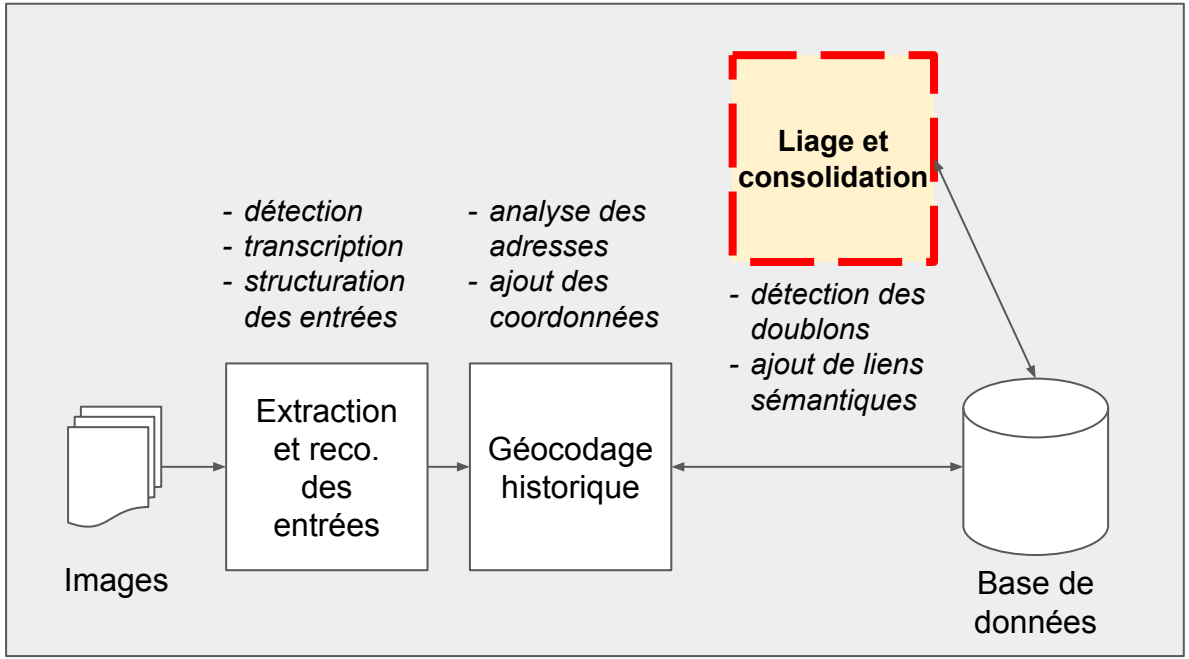
Didot 1842a - page 117

Didot 1843a - page 129

Didot 1844a - pages 125-126

Vue d'ensemble du processus

Définition du corpus

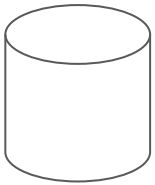


Analyse des données

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

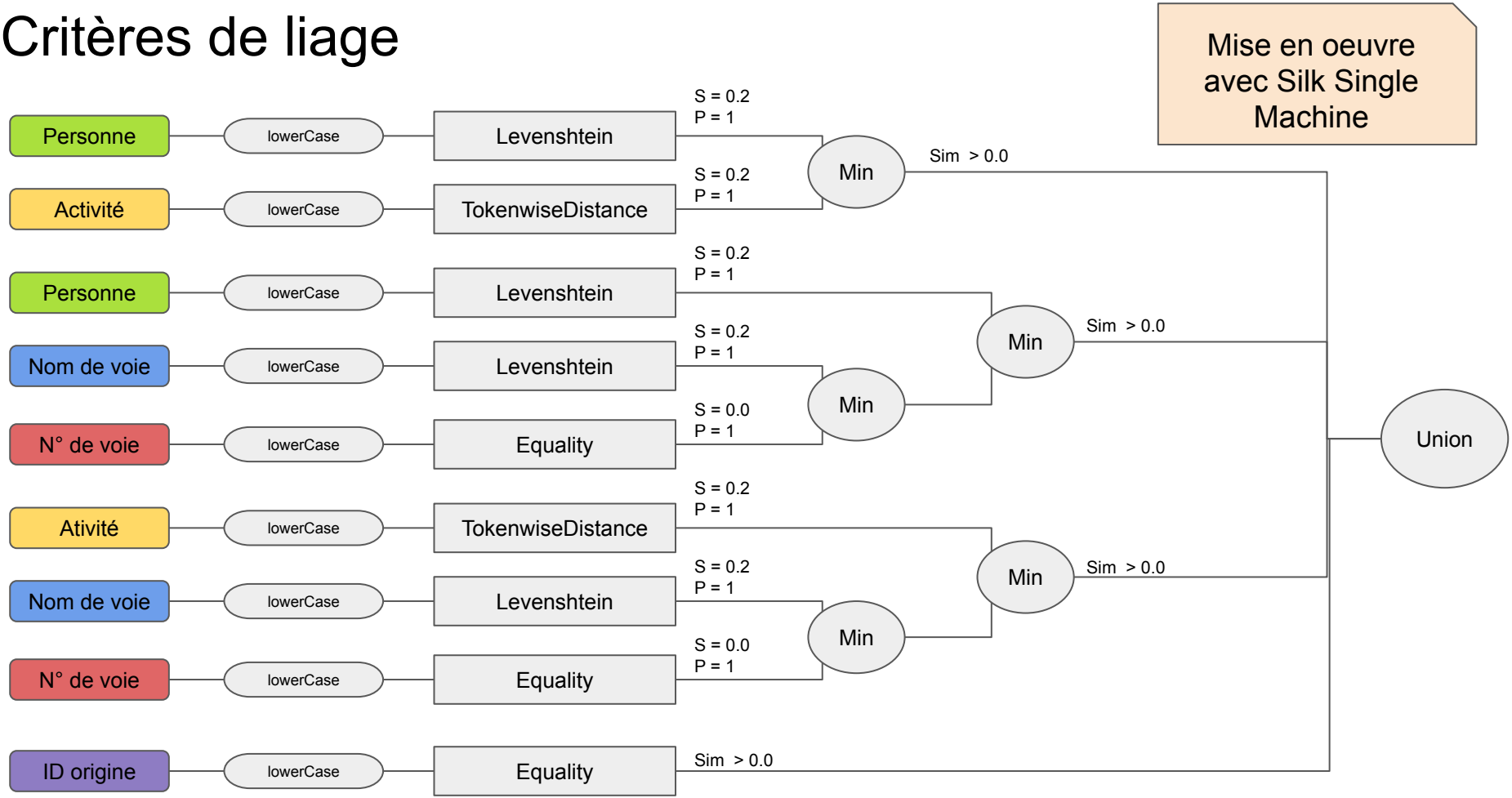


Base de données
22 743 928 entrées extraites

Mots-clés:
%atlas% and %cart%
%cart% and %geo%
%cart% and %géo%
%cart% and %marin%
%plan% and %topograph%
%cart% and %topograph%

1403 entrées d'annuaires

Critères de liage



Résultats du liage des entrées

Exemple des graveurs et marchands de cartes

Méthode numérique

20 362 liens

Paramétrage complexe :
identifier les seuils de tolérance pertinents

Comparaison adaptée aux chaînes de caractères résultant de l'OCR

+

Raisonnement

16 046 liens

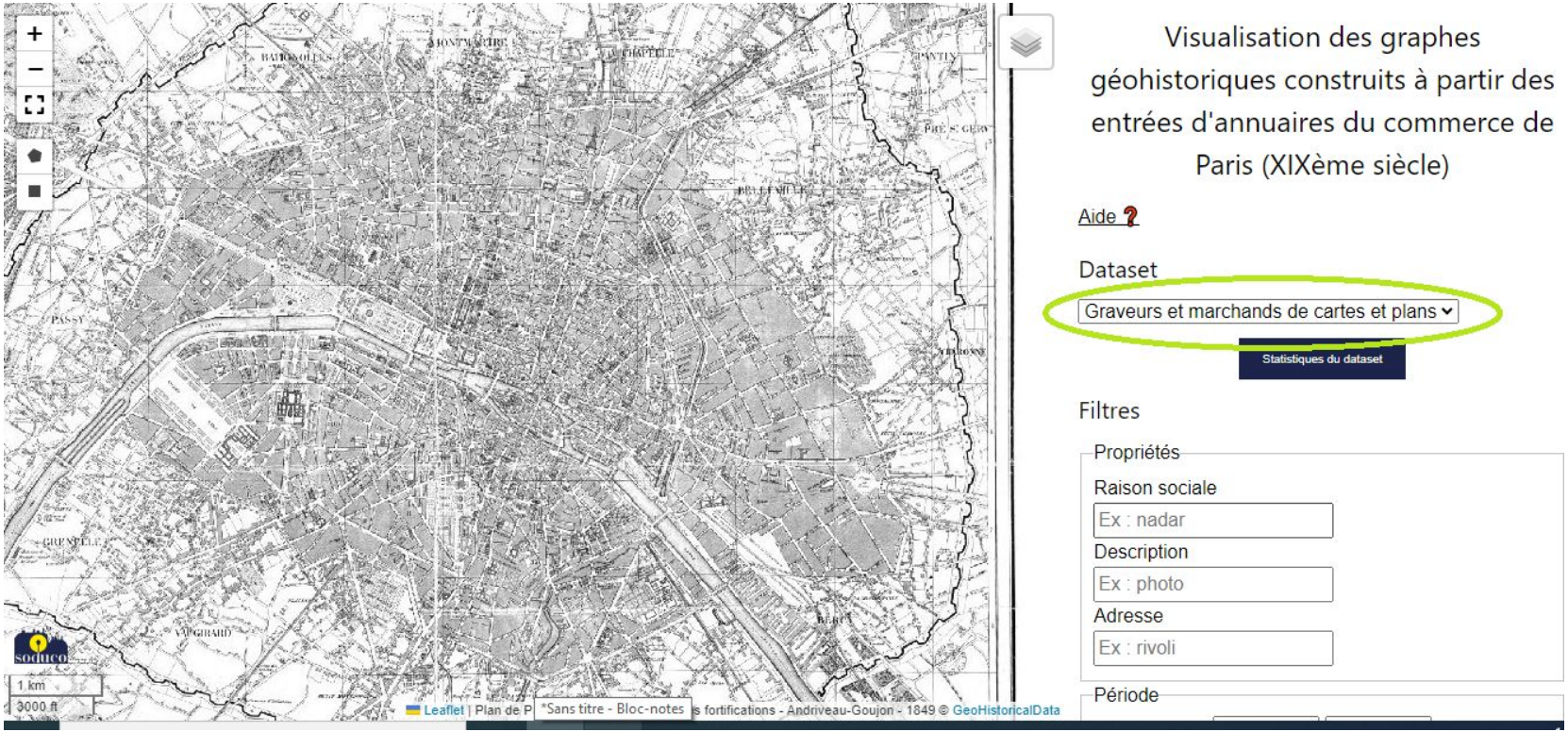
Propagation des liens owl:sameAs par transitivité

BILAN

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

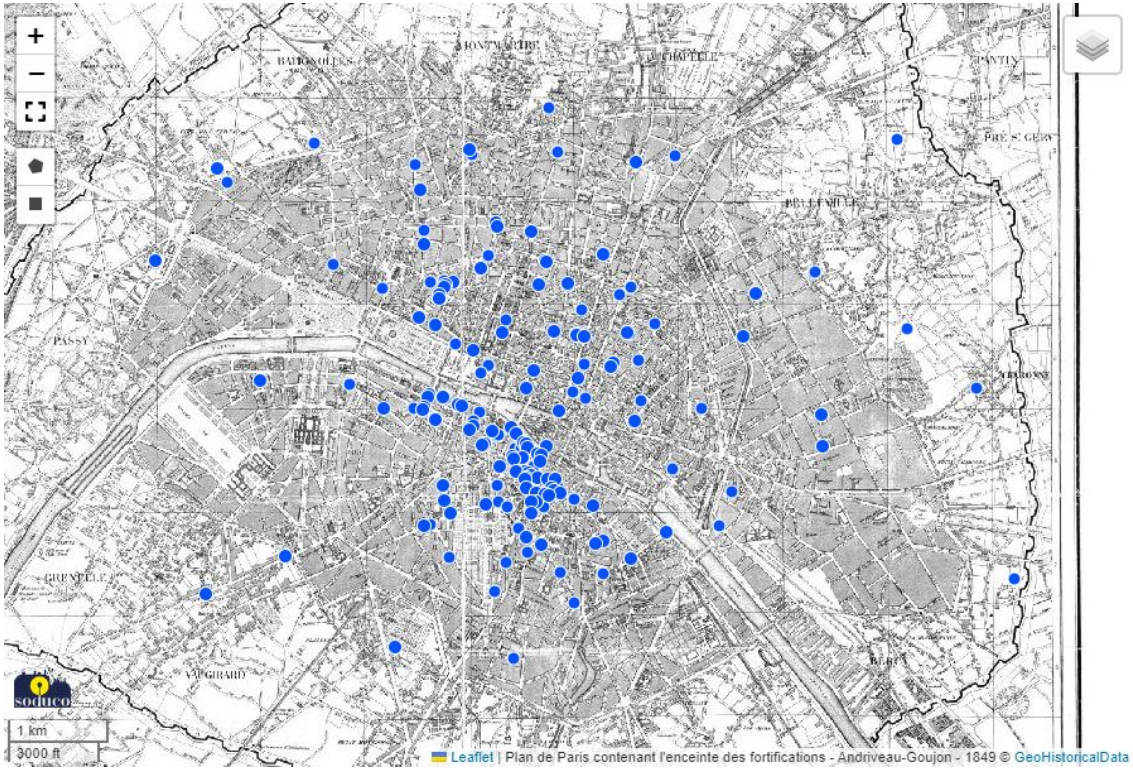
Graphe géohistorique : https://dir.geohistoricaldata.org/sparql

Visualisation et interprétation des résultats



https://soduco.geohistoricaldata.org/atelier_graphes_geohistoriques_annuaires/

Visualisation et interprétation des résultats



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/

Statistiques sur le graphe géohistorique

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

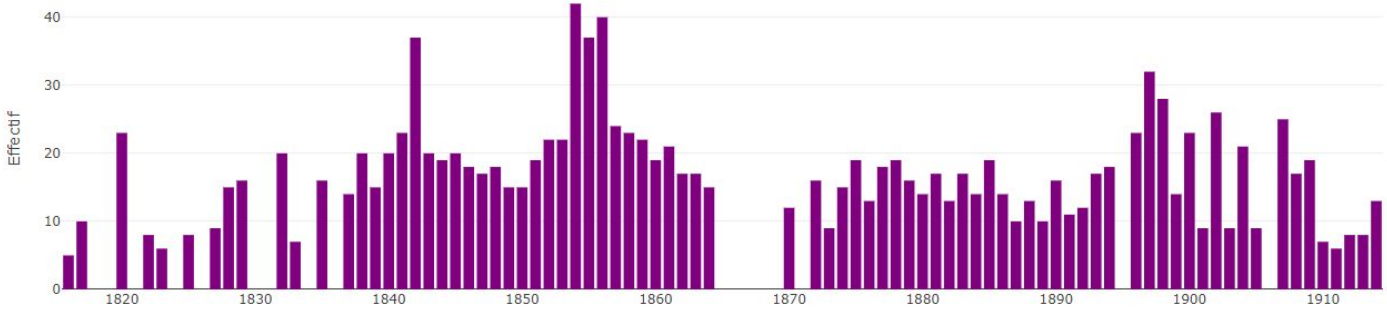
Nombre d'entrées d'annuaires
1403

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

Ex : photo

Adresse

Ex : rivoli

Période

1790 1915

1790 1821 1853 1884 1915

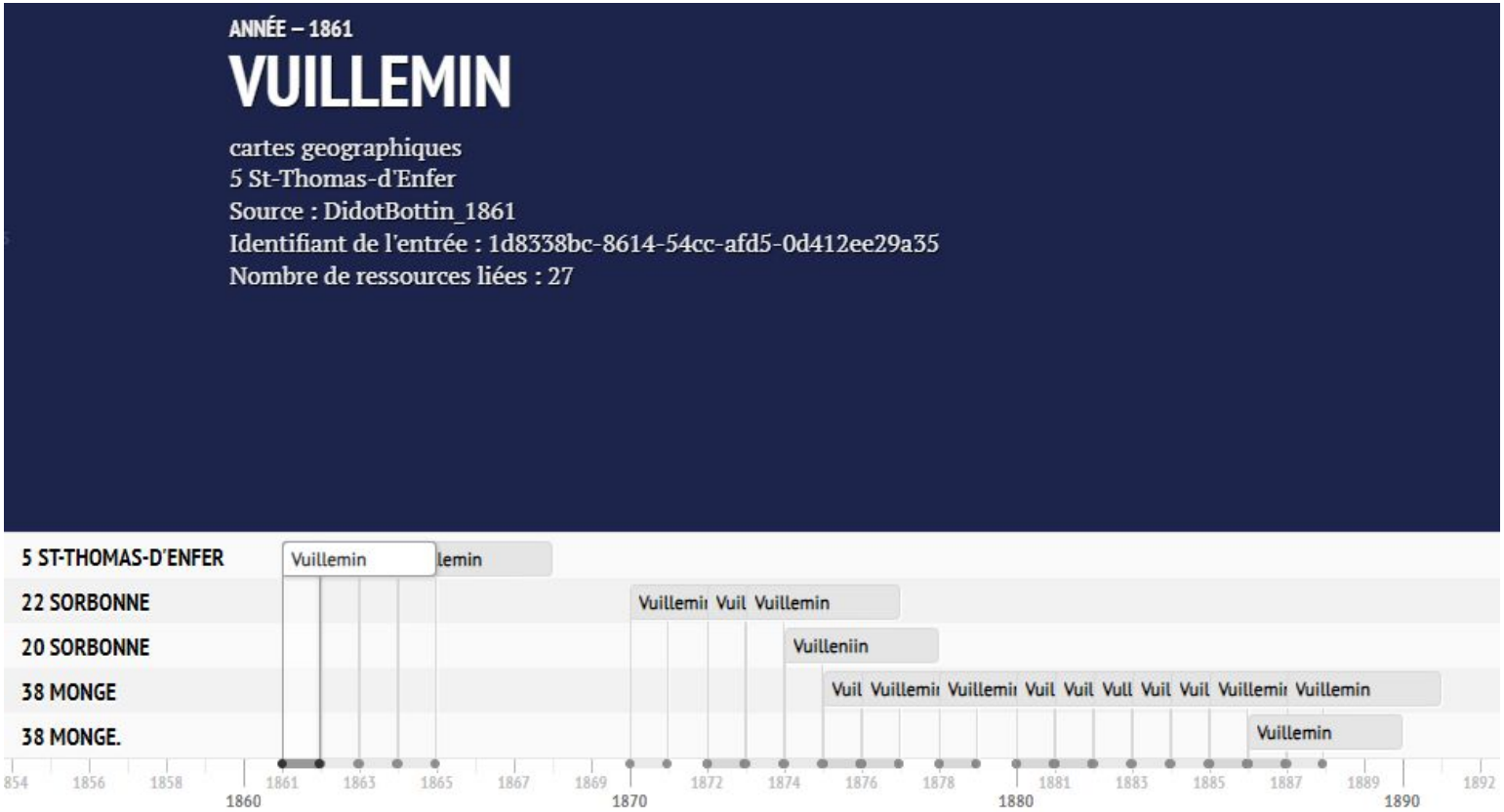
Le filtre temporel permet de faire varier l'affichage des points préalablement chargés sur la carte sans lancer une nouvelle recherche.
Données chargées pour la période **1790-1915**.

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

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

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

DOI 10.5281/zenodo.6394464

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

DOI 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_parisiens/

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/

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

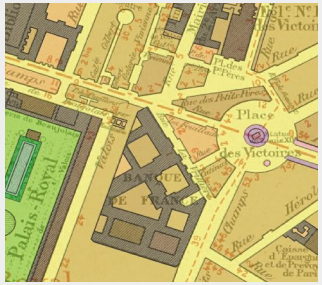
Reconnaissance → **Classification** → **Géo-référencement** → **Structuration**

Image → Géométries

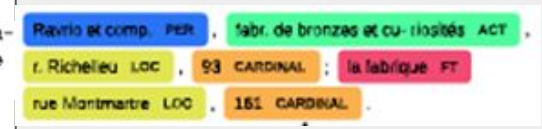
Scan → Texte

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

Géométries → Annotations

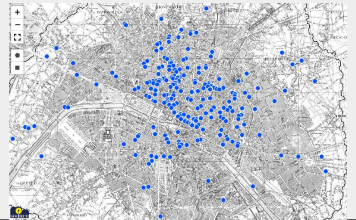


Texte → Entités nommées spatiales

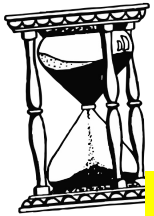


Données géométriques → Données géographiques

Entités nommées spatiales → Données géographiques

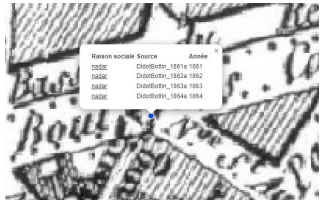


Données géographiques → Données géohistoriques

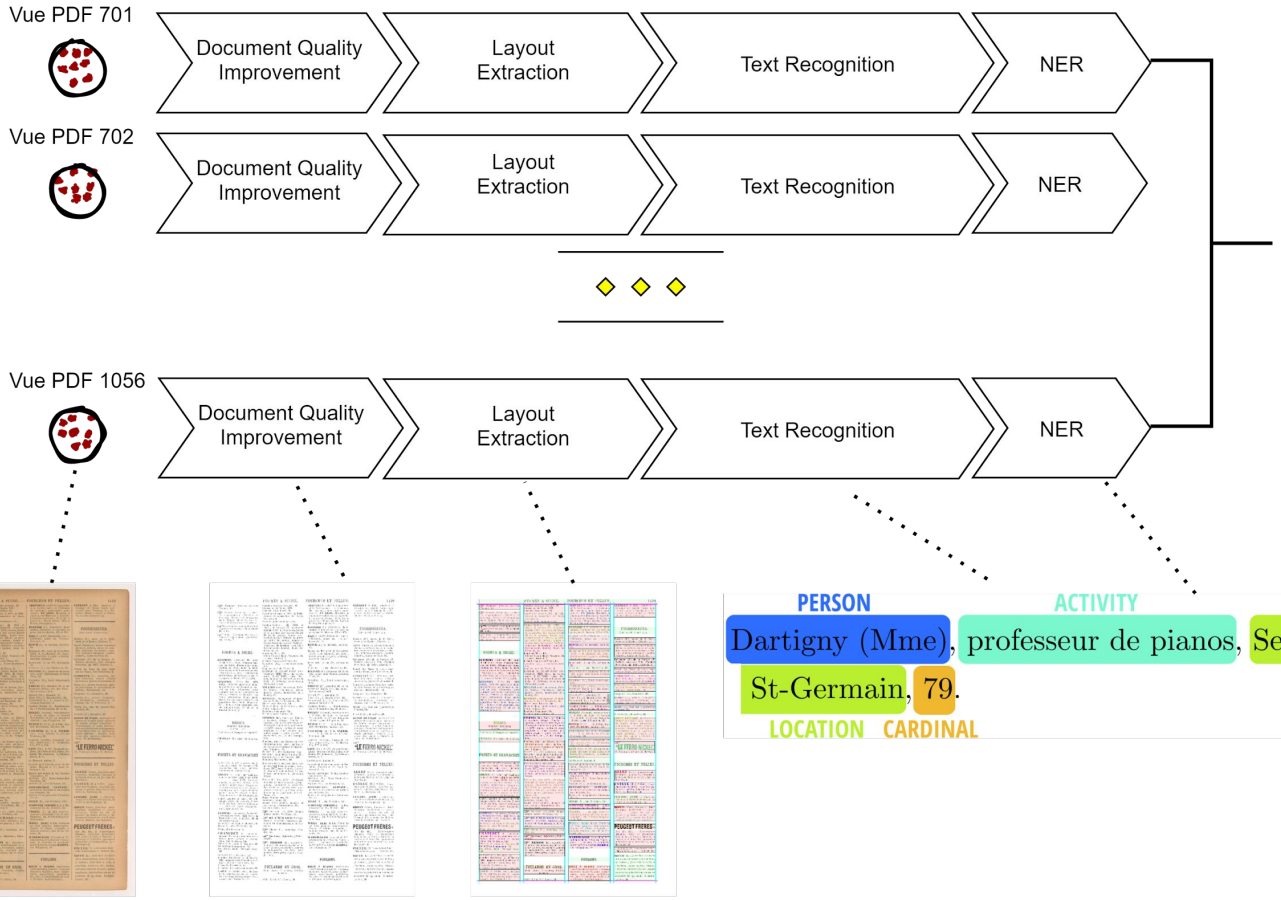


Aborder le liage ici ?

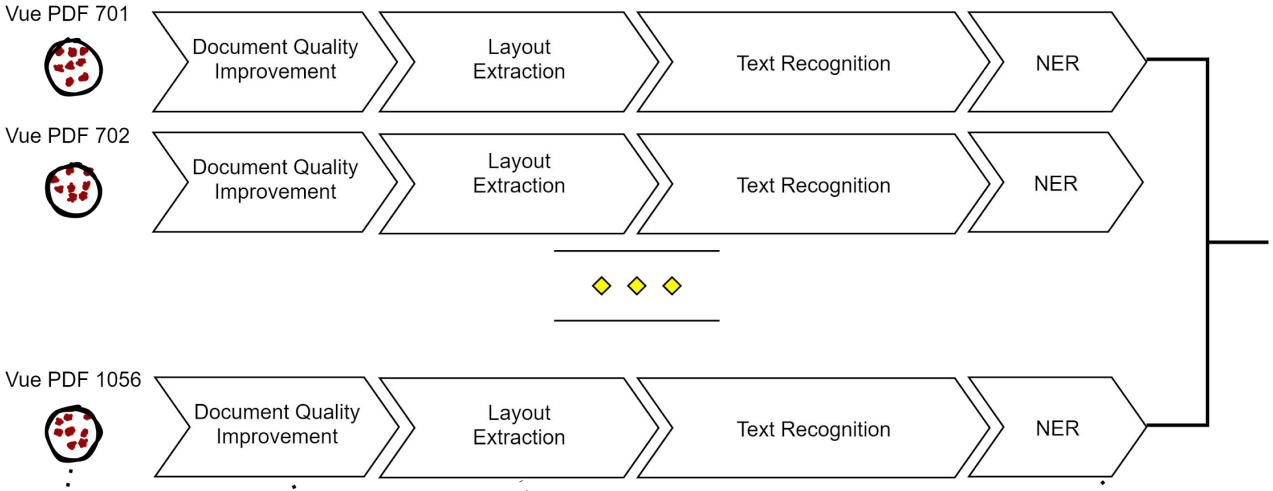
Données géographiques → Données géohistoriques



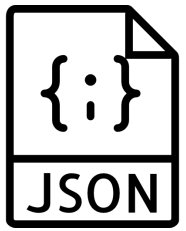
Chaîne de traitement automatique : de l'image aux infos sémantiques



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

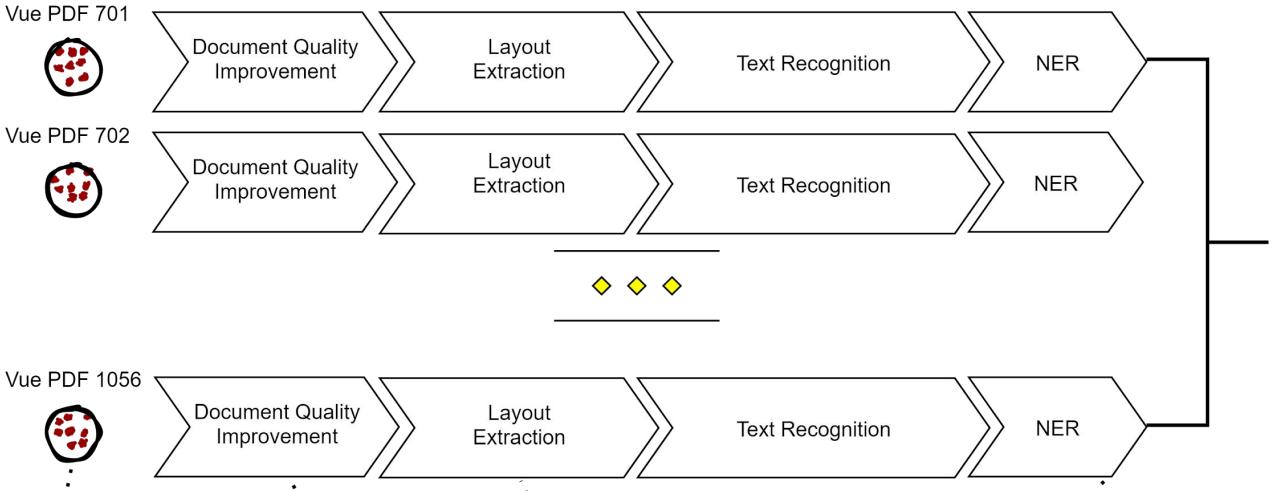


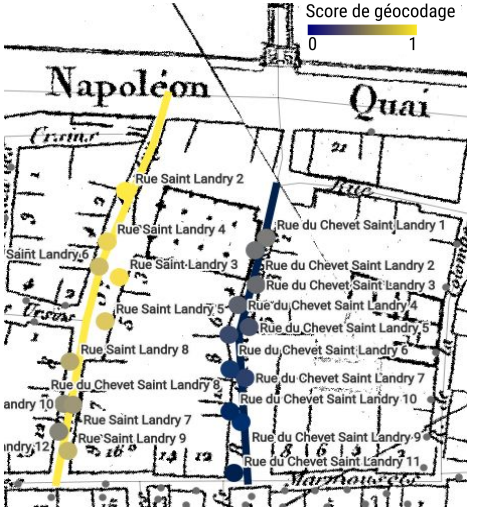
un JSON par vue



V2 data (2022-06): ~10 M d'entrées

Chaîne de traitement automatique : ajout des infos spatiales





PERSON ACTIVITY

Dartigny (Mme), professeur de pianos, Seine-St-Germain, 79.

LOCATION CARDINAL

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

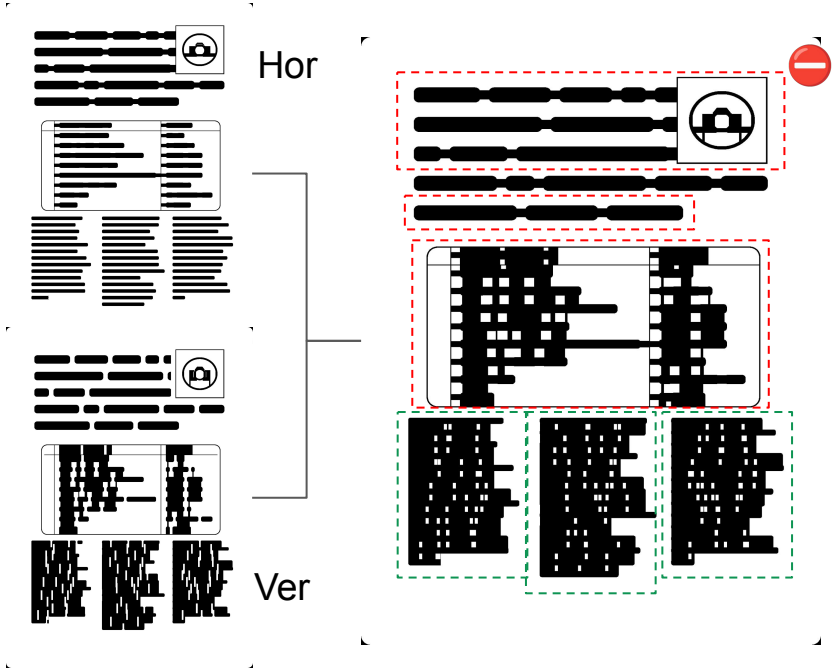
Free/open, accessible, indexé, reproductible
Dépôts GitHub, https://geohistoricaldata.org, jeux de données, publications…

Outils existants et ouverture en cours

Segmentation du canvas (pour les blocs)

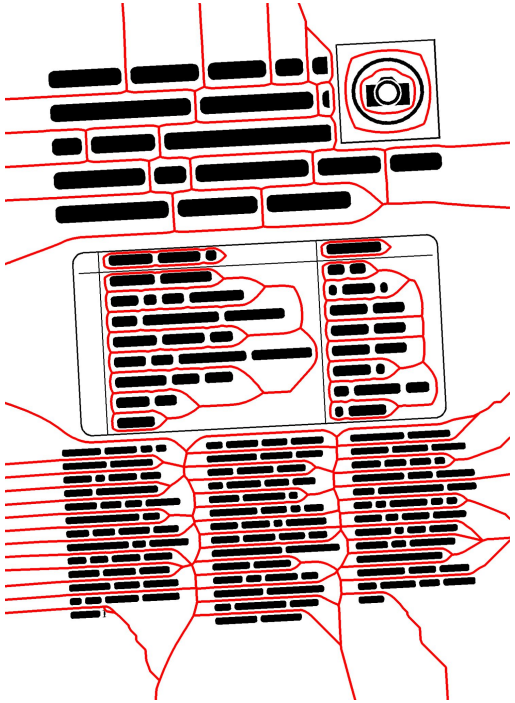
XY-Cut (Projection d'histogramme récu

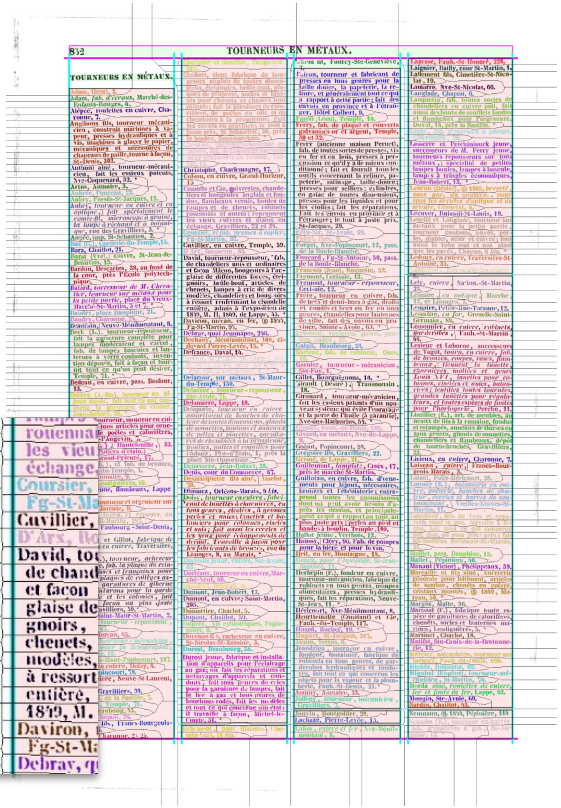
Smearing (ouverture/fermeture morphologique)

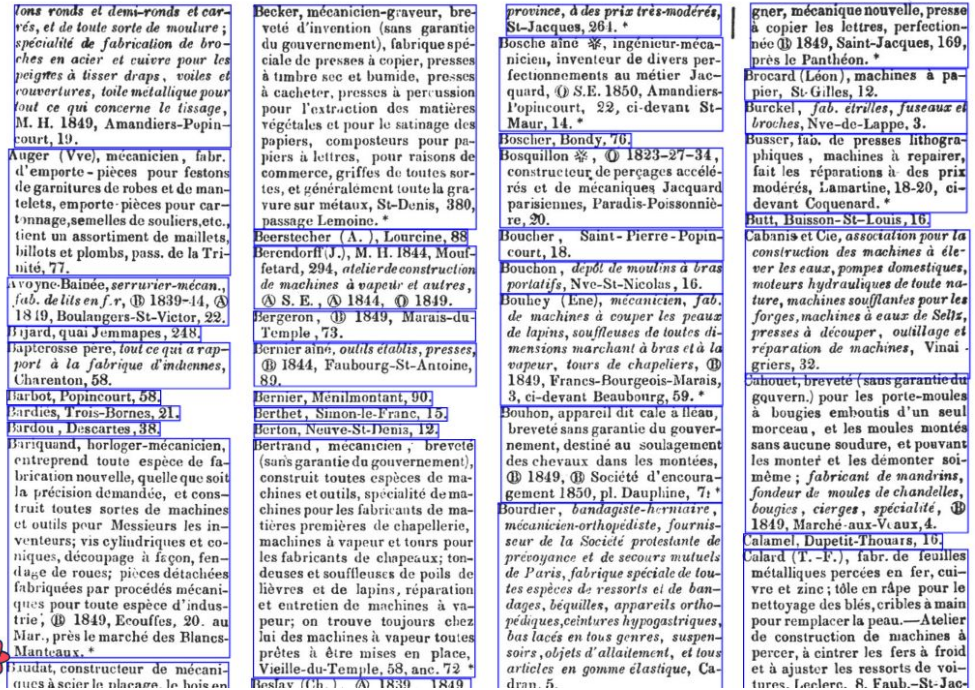


Segmentation du canvas (pour les lignes)

Ligne de partage des eaux







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

Auger (Vve), 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, billots et plombs, pass. de la Trinité, 77.

Avoyne-Bainée, *serrurier-mécan., fab. de lits en f. r*, Ⓑ 1839-14, Ⓐ 1849, Boulangers-St-Victor, 22.

Bijard, quai Jemmapes, 248.

Bapterosse père, *tout ce qui a rapport à la fabrique d'indiennes*, Charenton, 58.

Barbot, Popincourt, 58.

Bardies, Trois-Bornes, 21.

Bardou, Descartes, 38.

Bariquand, horloger-mécanicien, entreprend 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, Ecouffes, 20, au Mar., près le marché des Blancs-Manteaux. *

Baudat, constructeur de mécaniques à scier le placage, le bois en

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.), Lourcine, 88.

Berendorff (J.), M. H. 1844, Moufletard, 294, *atelier de construction de machines à vapeur et autres*, Ⓐ S. E., Ⓐ 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, 15.

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 soufflèuses 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 *

Beslay (Ch.), Ⓐ 1839 — 1849

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

Bosche aîné ✻, ingénieur-mécanicien, inventeur de divers perfectionnements au métier Jacquard, Ⓘ S.E. 1850, Amandiers-l'opincourt, 22, ci-devant St-Maur, 14. *

Boscher, Bondy, 76.

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.

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

Bouhon, appareil dit cale à l'eau, breveté sans garantie du gouvernement, destiné au soulagement des chevaux dans les montées, Ⓑ 1849, Ⓑ Société d'encouragement 1850, pl. Dauphine, 7 *

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.

gner, 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. étrilles, fuseaux et broches*, Nve-de-Lappe, 3.

Busser, fab. de presses lithographiques, machines à repairer, 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.

Cahouet, 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-Veaux, 4.

Calamel, Dupetit-Thouars, 16.

Calard (T.-F.), fabr. 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-

Directory ViewerError : Network Error EXPORT

Affichage Entrée

Edit inline

OCR NER

Cabanis et Cie PER , association pour la construction des machines à éle- 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 .

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

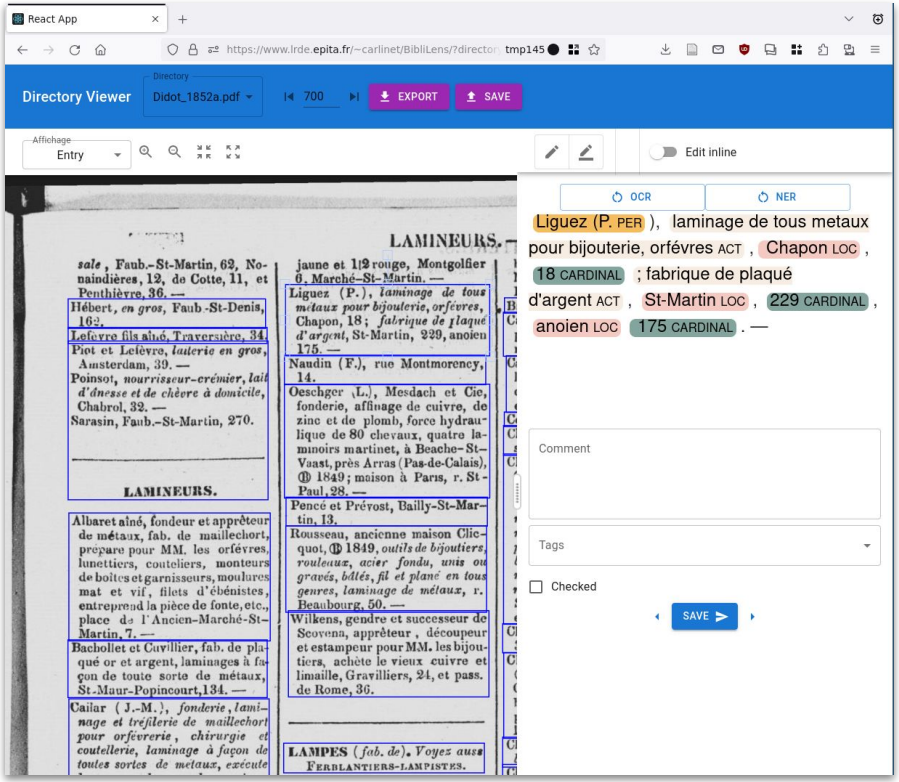
Développement de chaînes de traitement, pour permettre l'extraction de données pour des activités spécifiques (semaine "graphe géo-historique" en oct. 23)

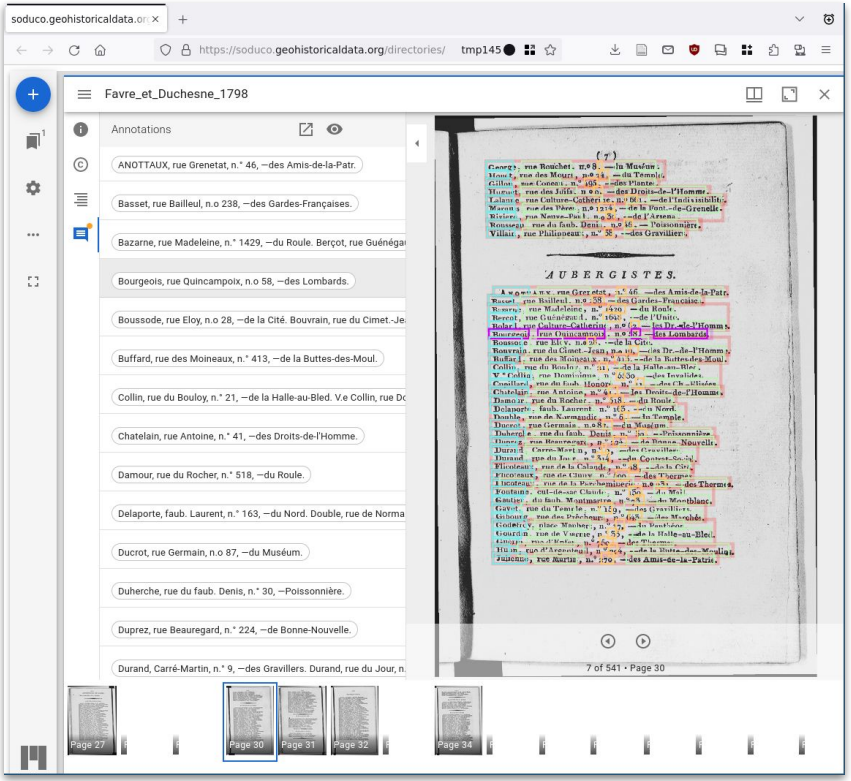
Croisement massif des données qui semble possible à court/moyen terme

Master puis thèse de Solenn Tual

- Suivi des photographes
- Transposition au cadastre napoléonien

Amélioration des outils d'annotation et de visualisation



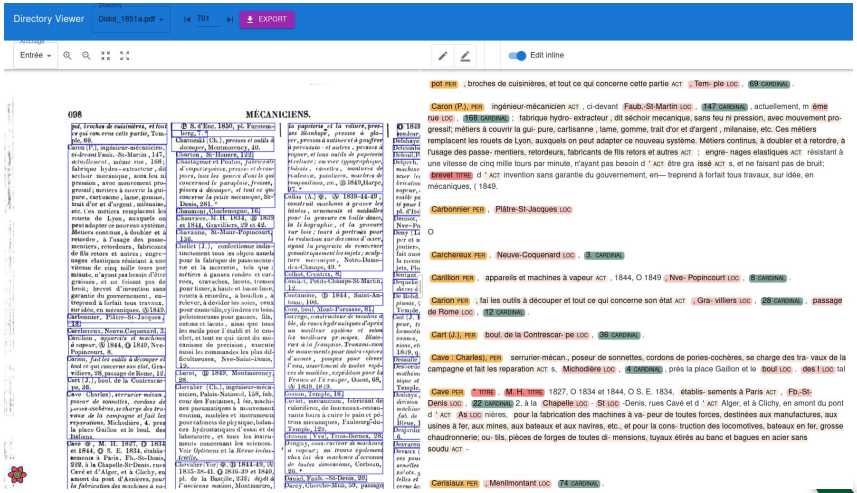


Construire un outil pour les historiens

Explorer  Visualiser  Annoter  Analyser 

API IIIF

E.g {BNF Gallica



Export structuré

{JSON}

TEI <XML />

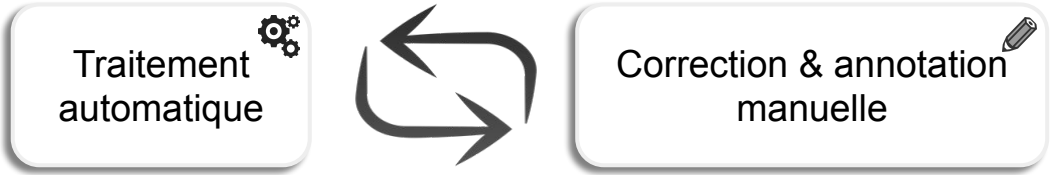
Assister le travail humain avec des outils automatiques

API IIIF
E.g {BNF Gallica

Export structuré
JSON / TEI XML

Interopérabilité

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



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

Un géocodeur historique sensible aux temps valides

Adresse
référence spatiale indirecte
r. St. Landry Cité, 2

Année cible
1838

Quadruplets
rues & adresses

Géocodeur

1. **Analyse syntaxique** *(parsing)*

<rue>	<quartier>	<numéro>
r. St. Landry	Cité,	2

2. **Recherche documentaire** *(collect,filter & score)*
3. **Tri temporel** *(reranking)*

Pelias-GH