Extraction automatique d'informations dans les annuaires commerciaux parisiens

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

Séminaire SoDUCo - BnF, 10 novembre 2022

(1) (2) (3)

SODUCO ANR-18-CE38-0013

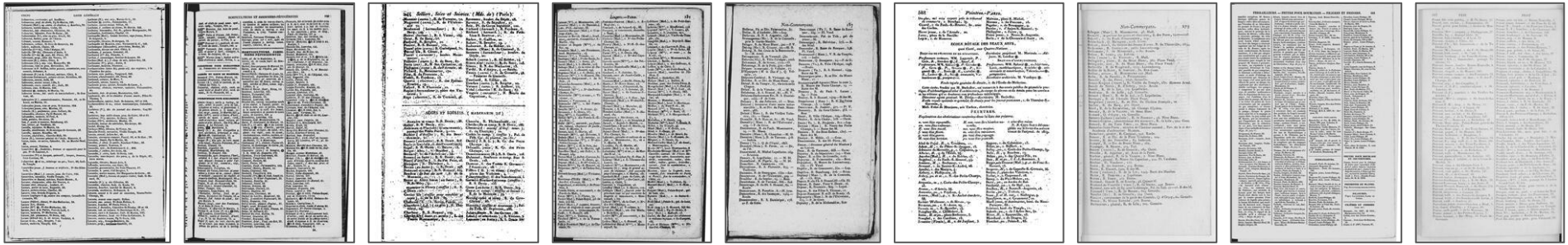
Introduction

Un corpus de **141** annuaires numérisés :

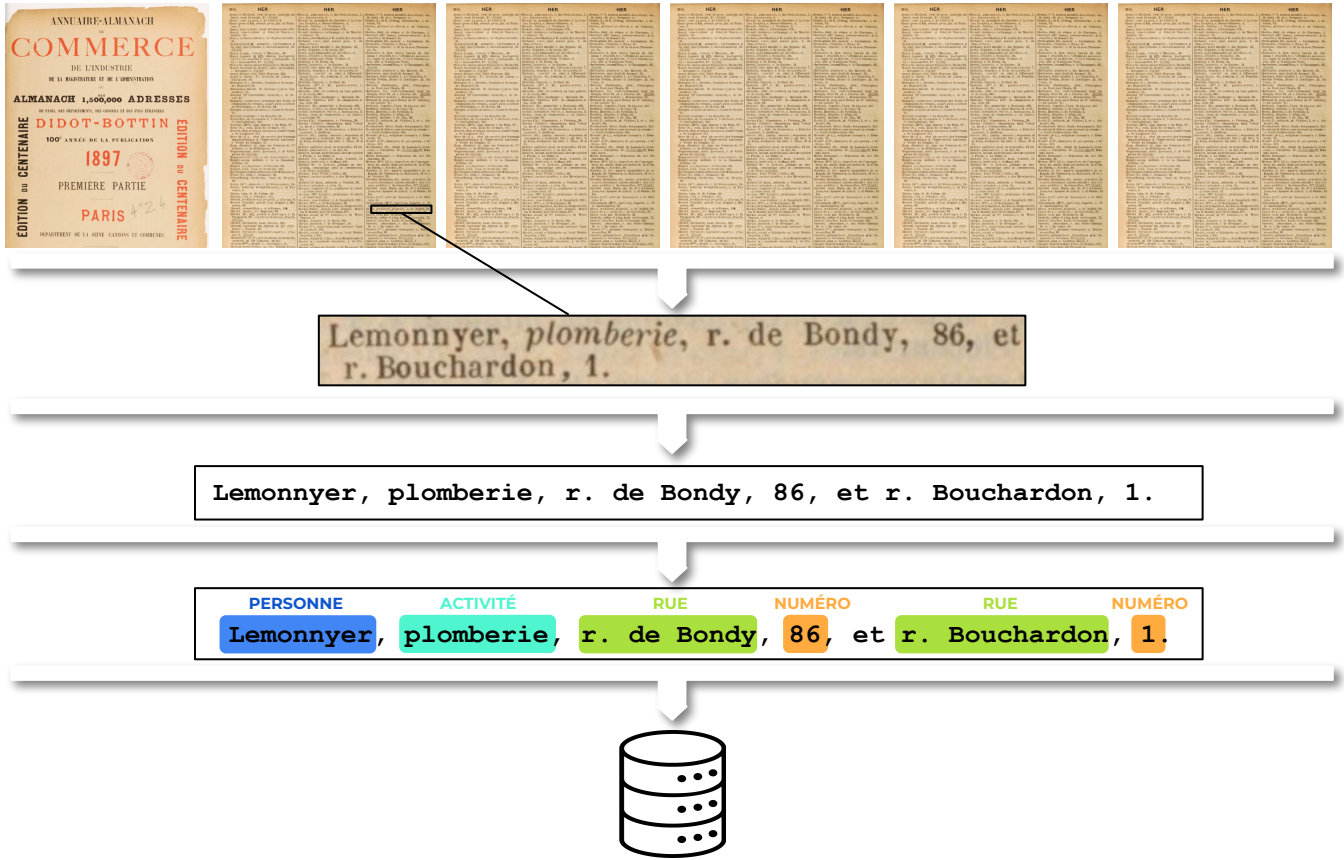
- Plusieurs centaines de milliers de pages,
- Plusieurs millions d'**entrées,** comportant le nom des commerces, leur type d'activité et leur localisation.

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

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



Des documents numérisés aux données structurées



Pages d'annuaires numérisées

Identification des entrées

Entrée d'annuaire

Transcription du texte

Texte numérique de l'entrée

Annotation du texte

Entités nommées

Structuration des entrées

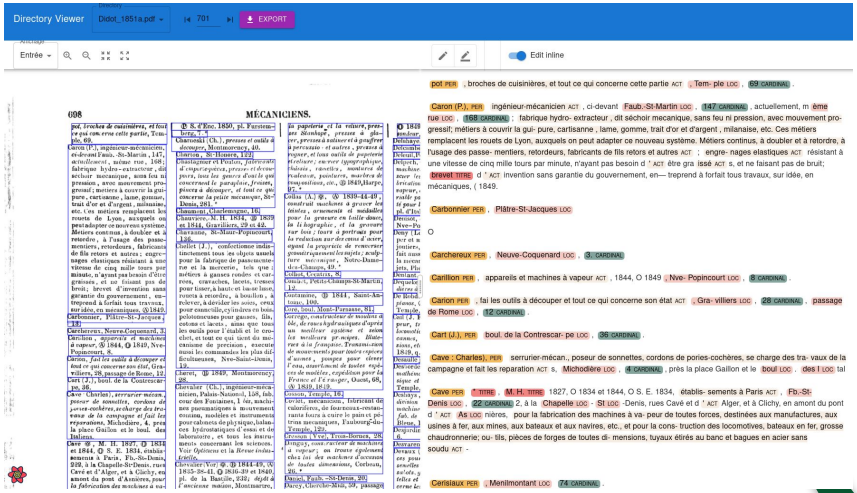
Base de données intégrée

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

Explorer | Visualiser | Annoter | Analyser

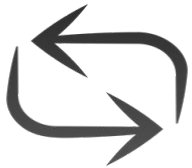
API IIIF
E.g {BNF Gallica

Export structuré
JSON / TEI XML

Interopérabilité

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

Traitement automatique

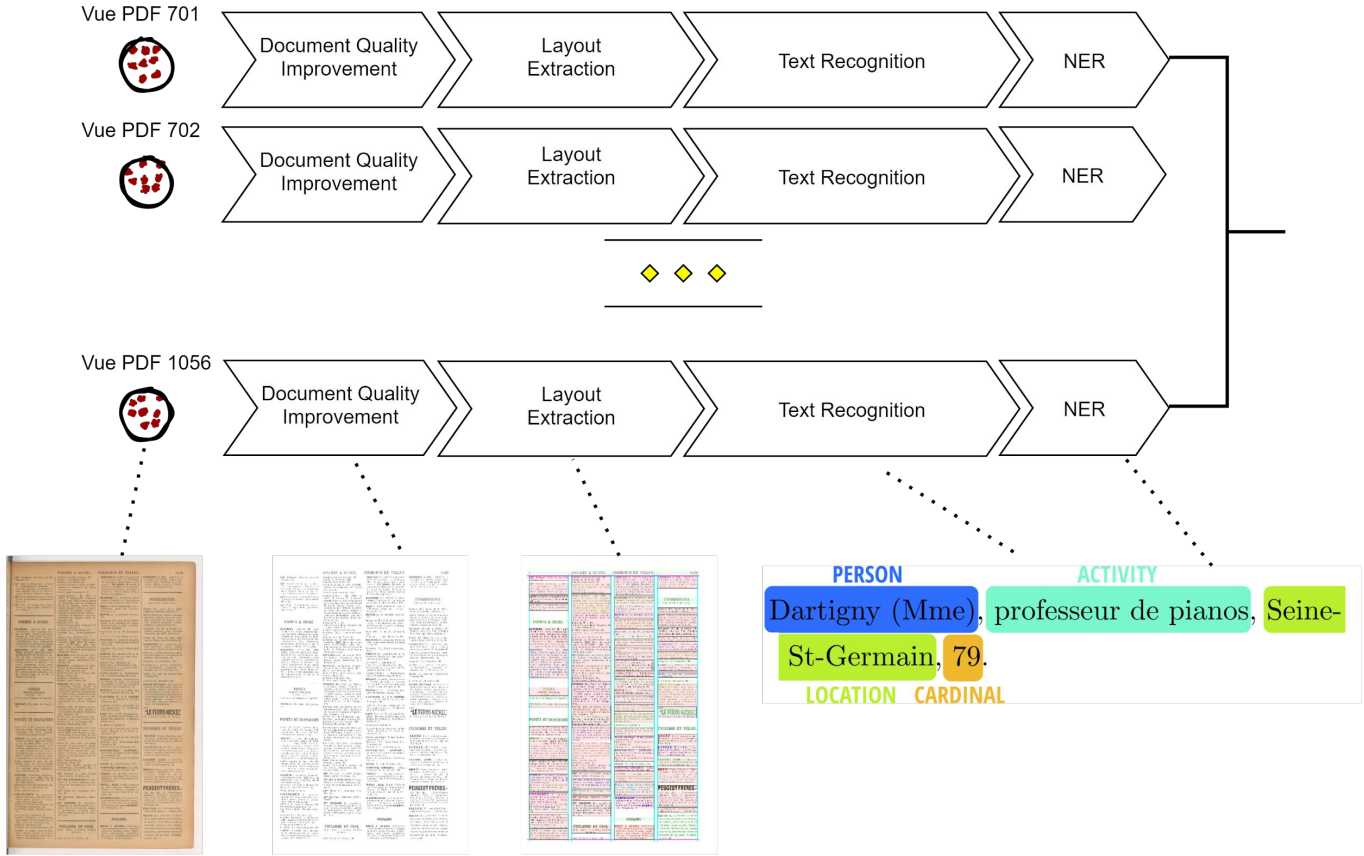


Correction & annotation manuelle

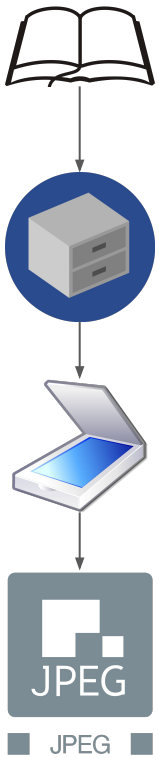
D'autres avantages:

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

Chaîne de traitement automatique



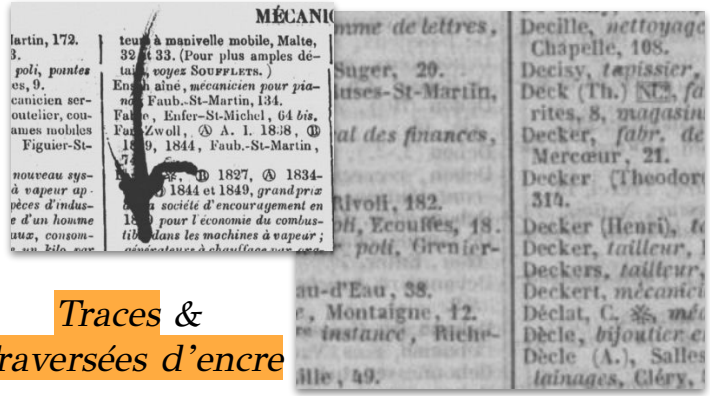
Une reconnaissance perturbée par différentes imperfections



Archivage

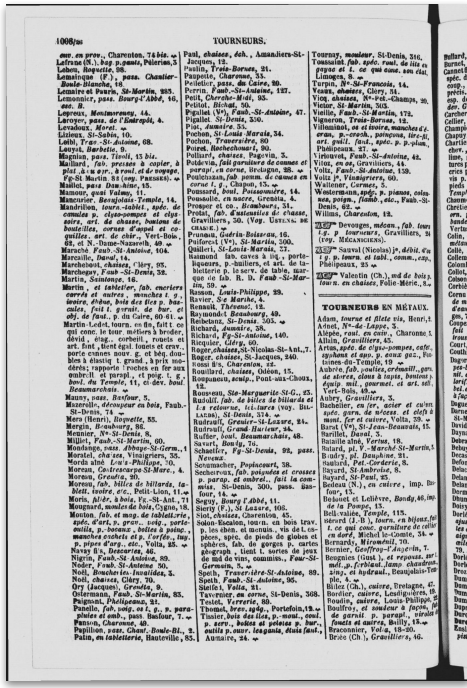
Numérisation

Compression



Perte des niveaux de gris & Artefacts de compression

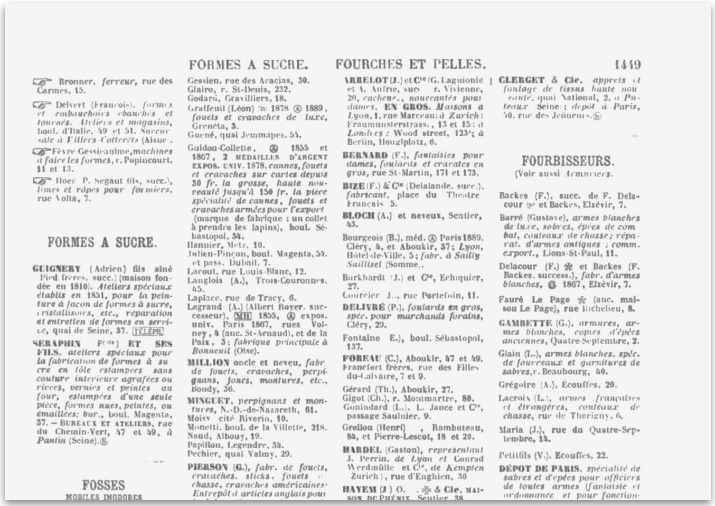
Inclinaison & Rotation & Courbure



Amélioration de la qualité des images

👉 Les OCRs fonctionnent mieux sur des images “propres“ :

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



Extraction de canvas / Segmentation

Méthode rapide basée XY-Cut pour les blocs

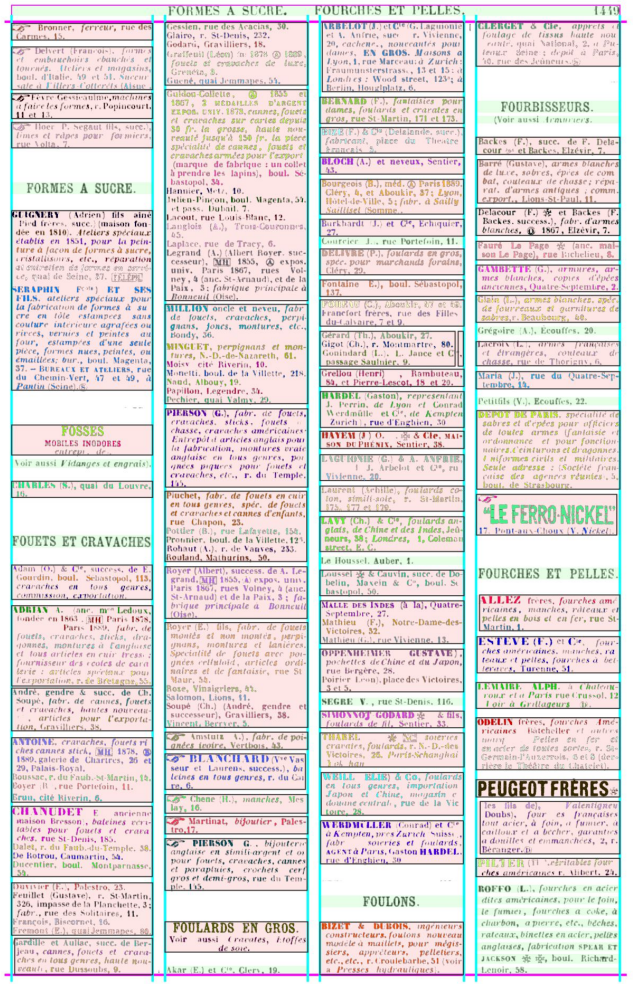
- Plusieurs niveaux hiérarchiques
- Gestion efficace des colonnes/blocs qui exploite la redondance visuelle des pages

Méthode basée Watershed pour les lignes / entrées

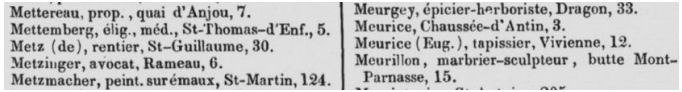
- L'indentation des lignes / espaces de fin de lignes permettent de déterminer un début d'entrée

Limites

- Sensible au bruit, d'où le nettoyage en amont
- Classification simultanée des régions
- Information "logique" intégrée (gestion du multi-page...)



Reconnaissance du texte (OCR)


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

Analyse de la mise en page

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

OCR

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

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

- ➔ *Kraken*
- ➔ *Tesseract* ⭐
- ➔ *Pero OCR* ⭐⭐

Reconnaître et structurer les informations des entrées

Composition des entrées:

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

L'ordre des informations peut varier selon le type d'index ou l'éditeur.
Seules les informations en gras sont systématiquement présentes.
Les informations suivies d'une * peuvent apparaître plusieurs fois par entrée.
Les étapes précédentes peuvent avoir produit un texte bruité : entrées mal délimitées, caractères mal reconnus, manquants, etc.

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

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

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

N. Abadie, E. Carlinet, Joseph Chazalon, B. Duménieu. A Benchmark of Named Entity Recognition Approaches in Historical Documents: Application to 19th Century French Directories. Document Analysis Systems. DAS 2022., May 2022, La Rochelle, France.
https://github.com/soduco/paper-ner-bench-das22

Modèles NER évalués

Nécessaire pour traiter des entités non standards

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

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

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

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

Quantité d'annotations et qualité des résultats

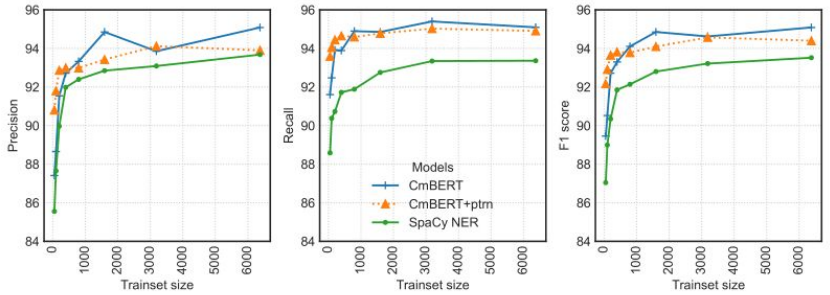
2. A-t-on besoin de beaucoup de données d'entraînement ?

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

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

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

	Training examples	49	99	199	398	796	1593	3186	6373
	%	0.8	1.6	3.1	6.2	12.5	25.0	50.0	100.0
F1 score	CmBERT	89.5	90.5	92.7	93.3	**94.1**	**94.9**	**94.6**	**95.1**
	CmBERT-ptrn	**92.2**	**92.9**	**93.6**	**93.8**	93.8	94.1	**94.6**	94.4
	SpaCy NER	87.0	89.0	90.3	91.9	92.1	92.8	93.2	93.5



F1-scores moyens obtenus pour 5 entraînements + tests

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

Modèles testés :
- CamemBERT
- CamemBERT pré-entraîné

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

Sorties — Prédictions :
- prédictions NER de camemBERT
  Avec et sans pré-entraînement

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

Mesures : score F1

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

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

Extrait du jeu de test de référence:

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

Extrait du jeu de test avec annotations projetées sur du texte bruité :
Dulay ***PER***, chandronnier ***ACT***, +. du Pont-anx-Cars ***LOC*** Ge 7 ***CARDINAL*** Fe ÊR one

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

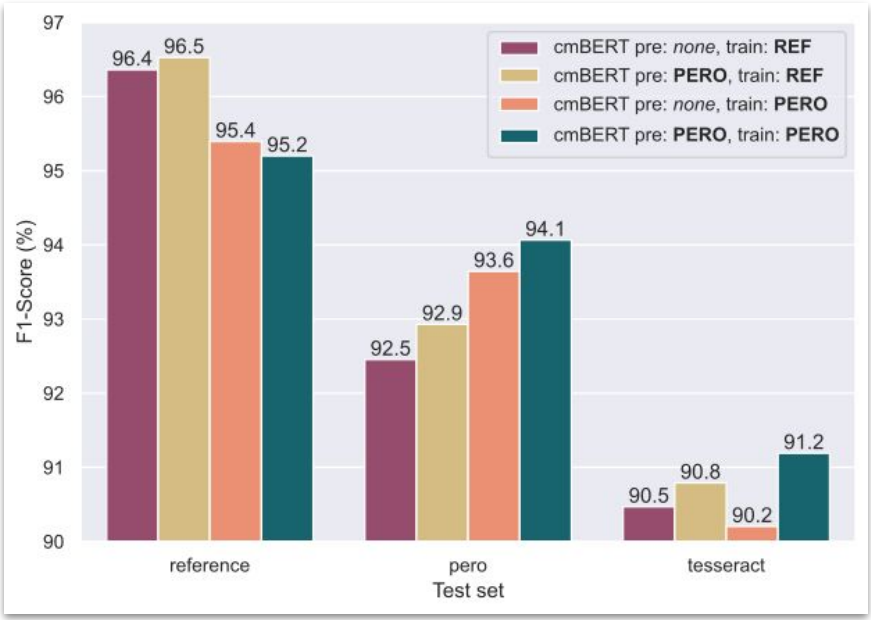
3 bis. Les modèles de NER à base de réseaux de neurones profonds sont-ils adaptés pour traiter des **transcriptions OCR bruitées** ?

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

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

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

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



F1-scores moyens obtenus pour 5 entraînements + tests

Démonstration

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-Popincourt, 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, Frances-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 voi-tures. 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 .

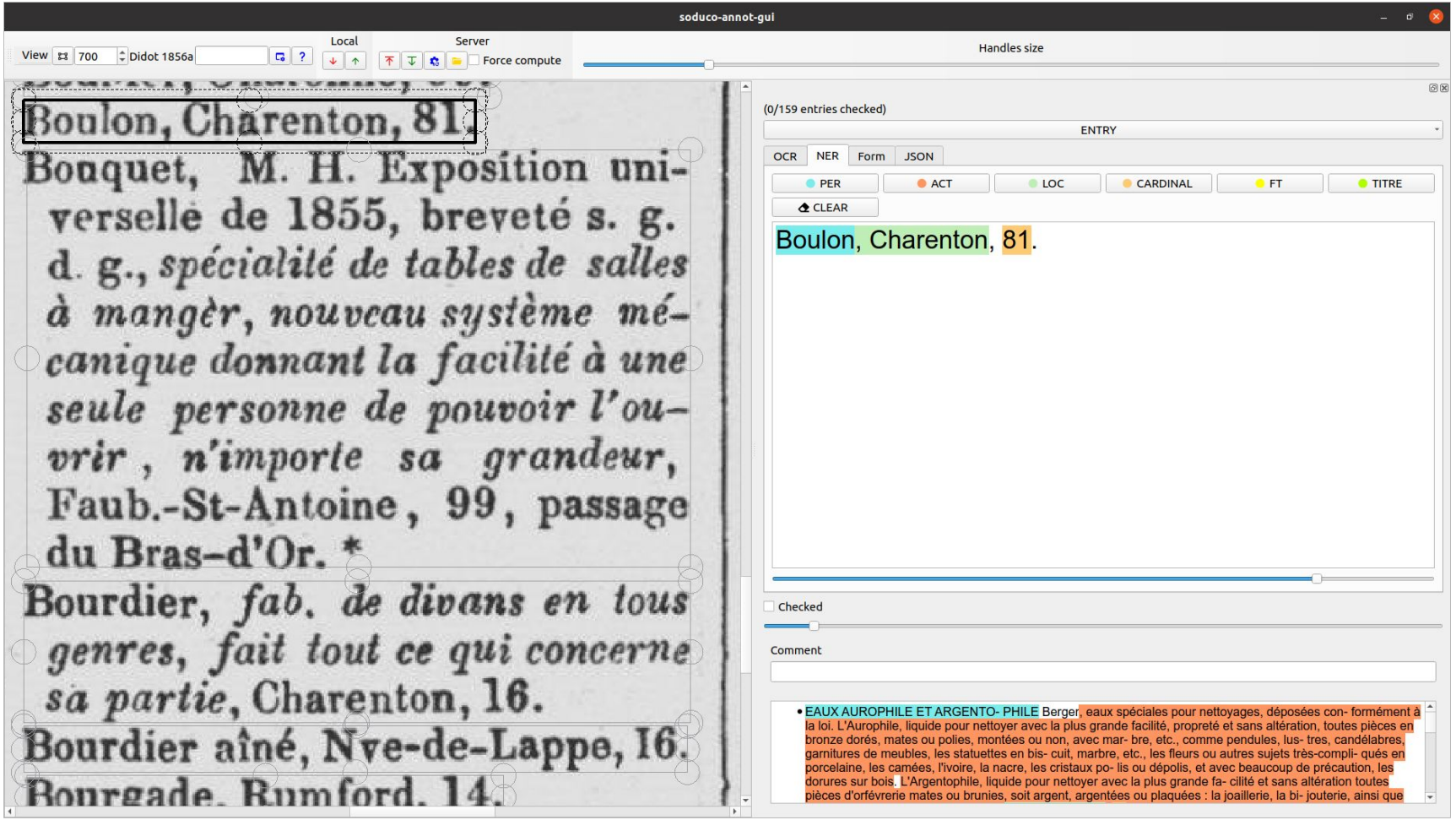
Comment

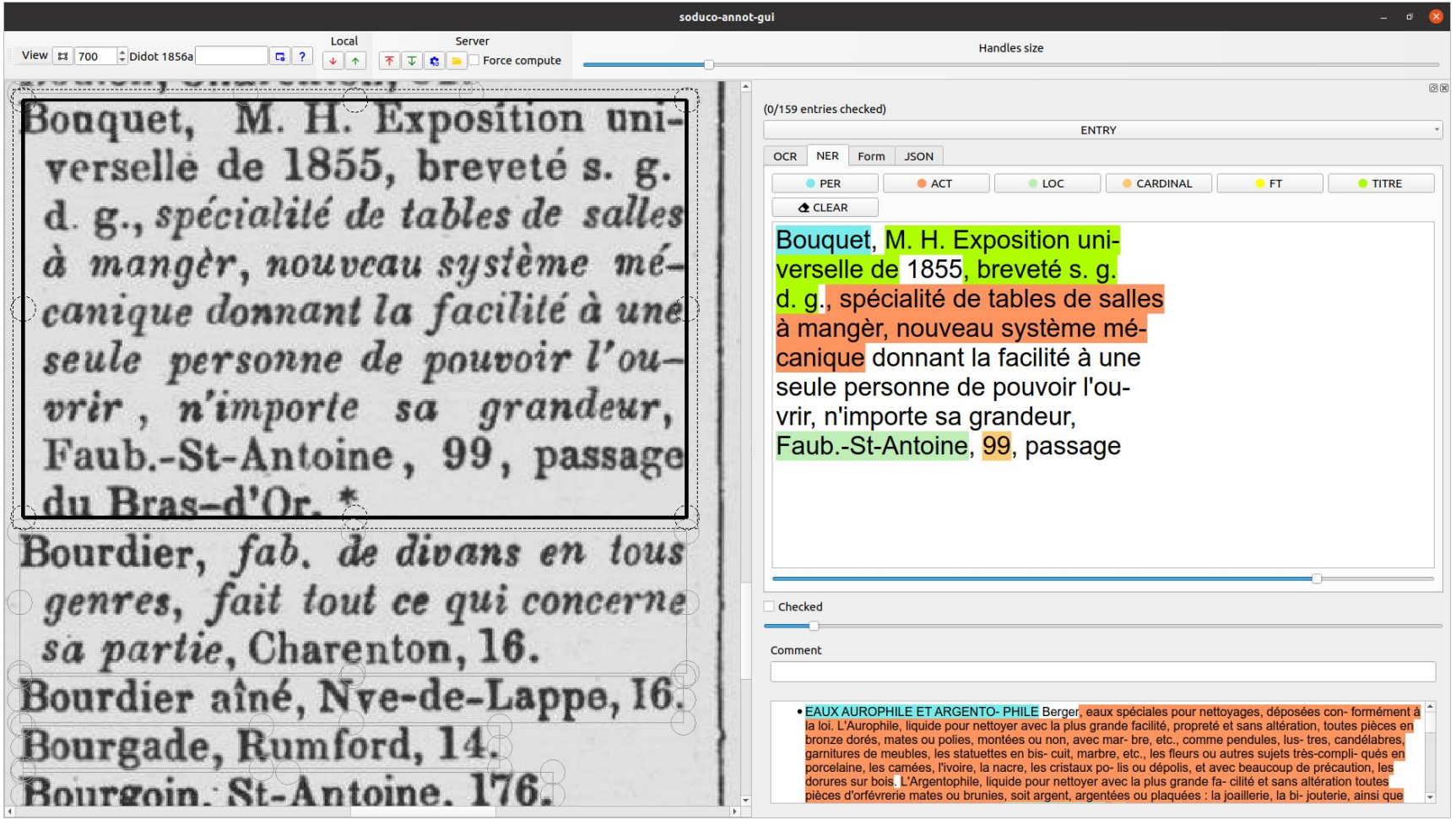
Tags

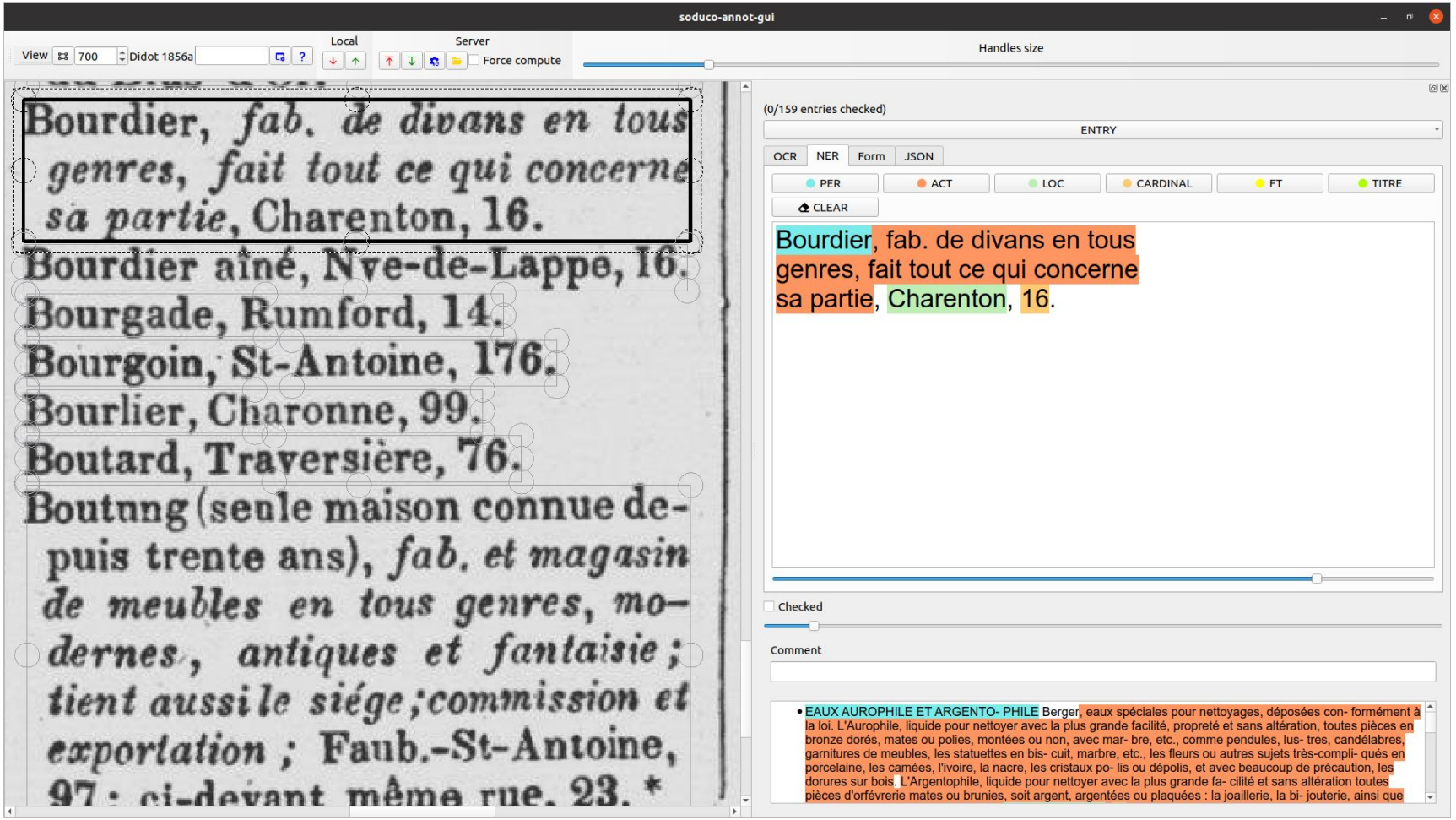
Checked

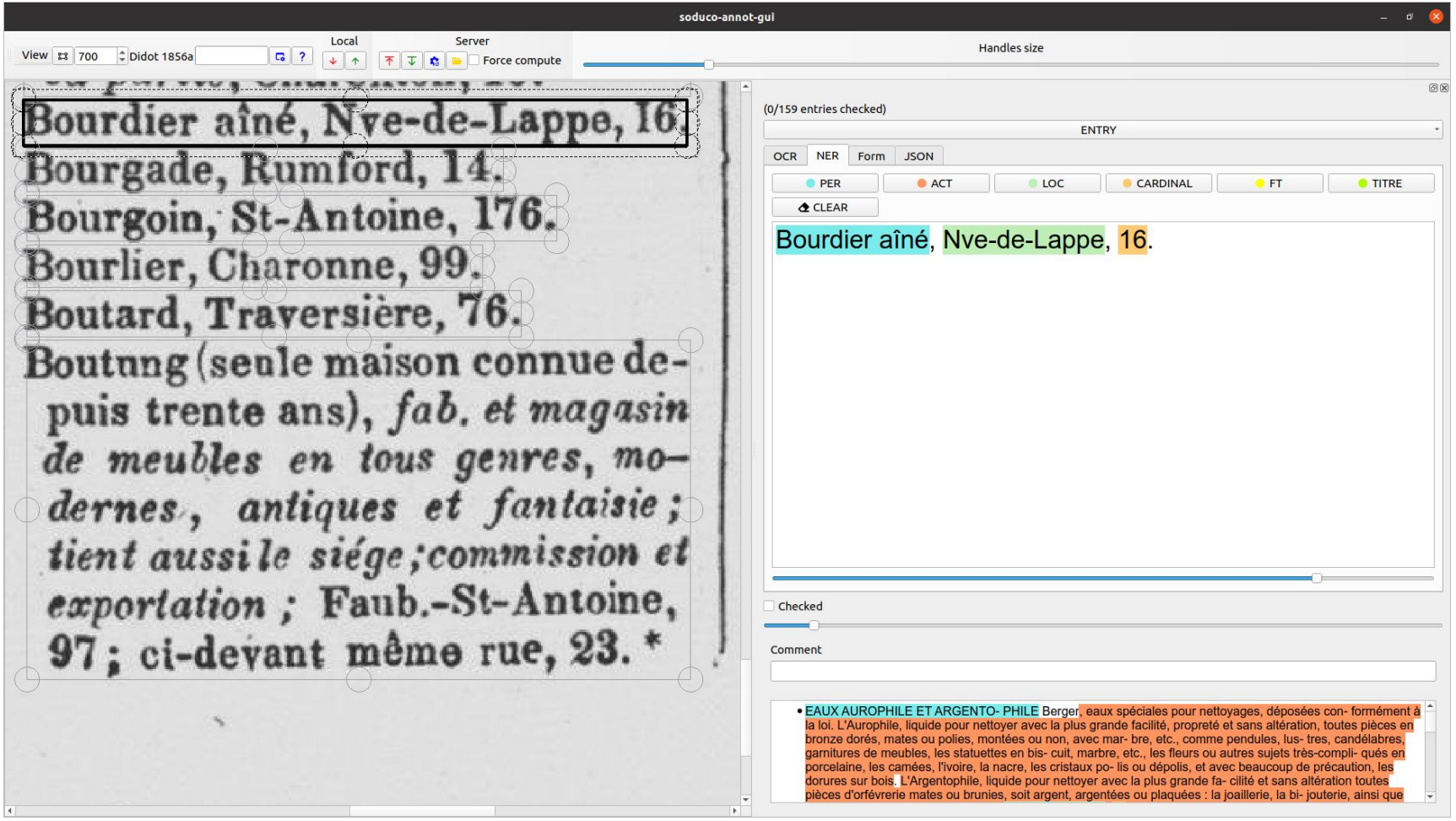
SAVE

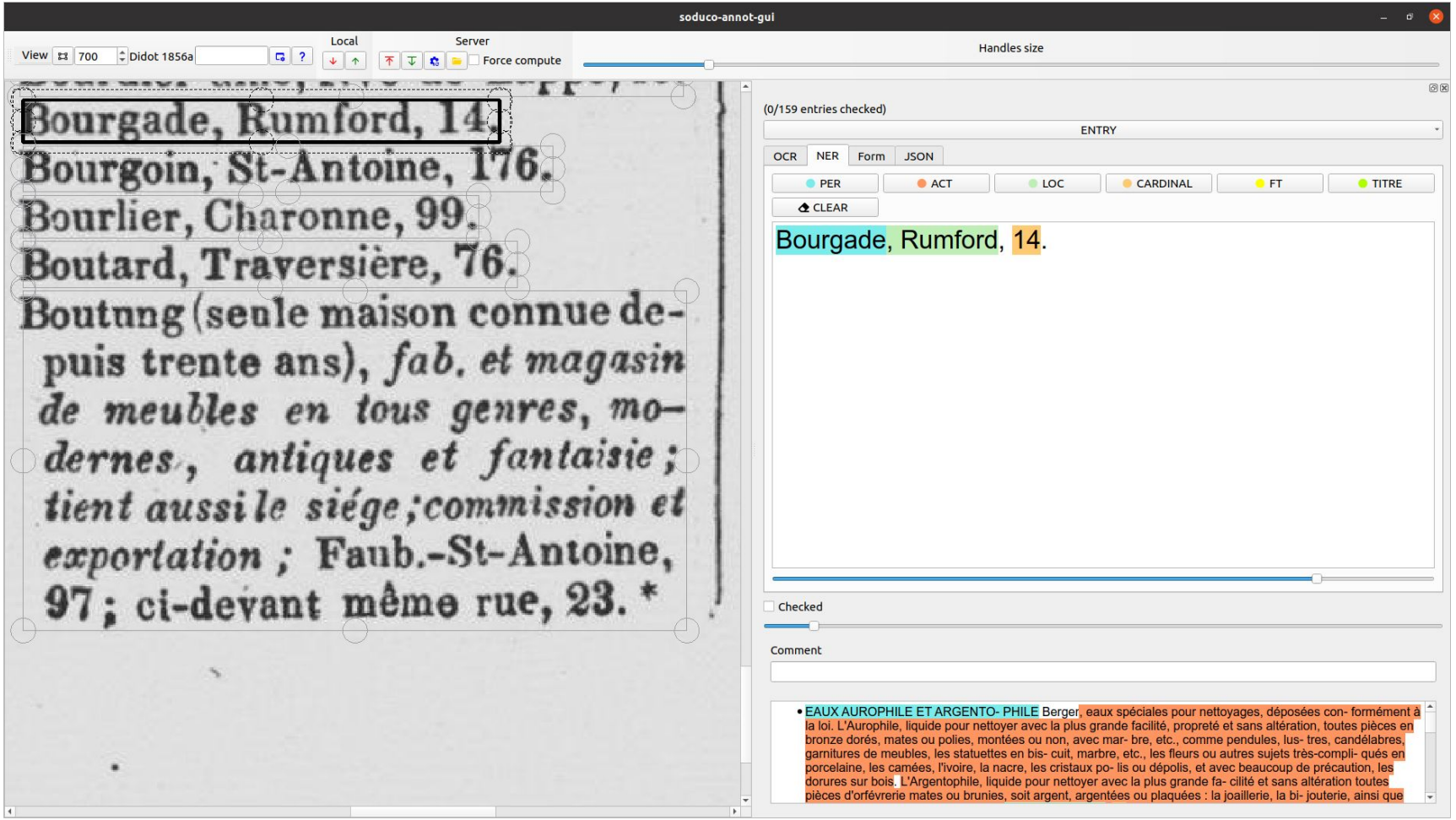
Quelques résultats qualitatifs
1. Cas positifs (ancienne interface)

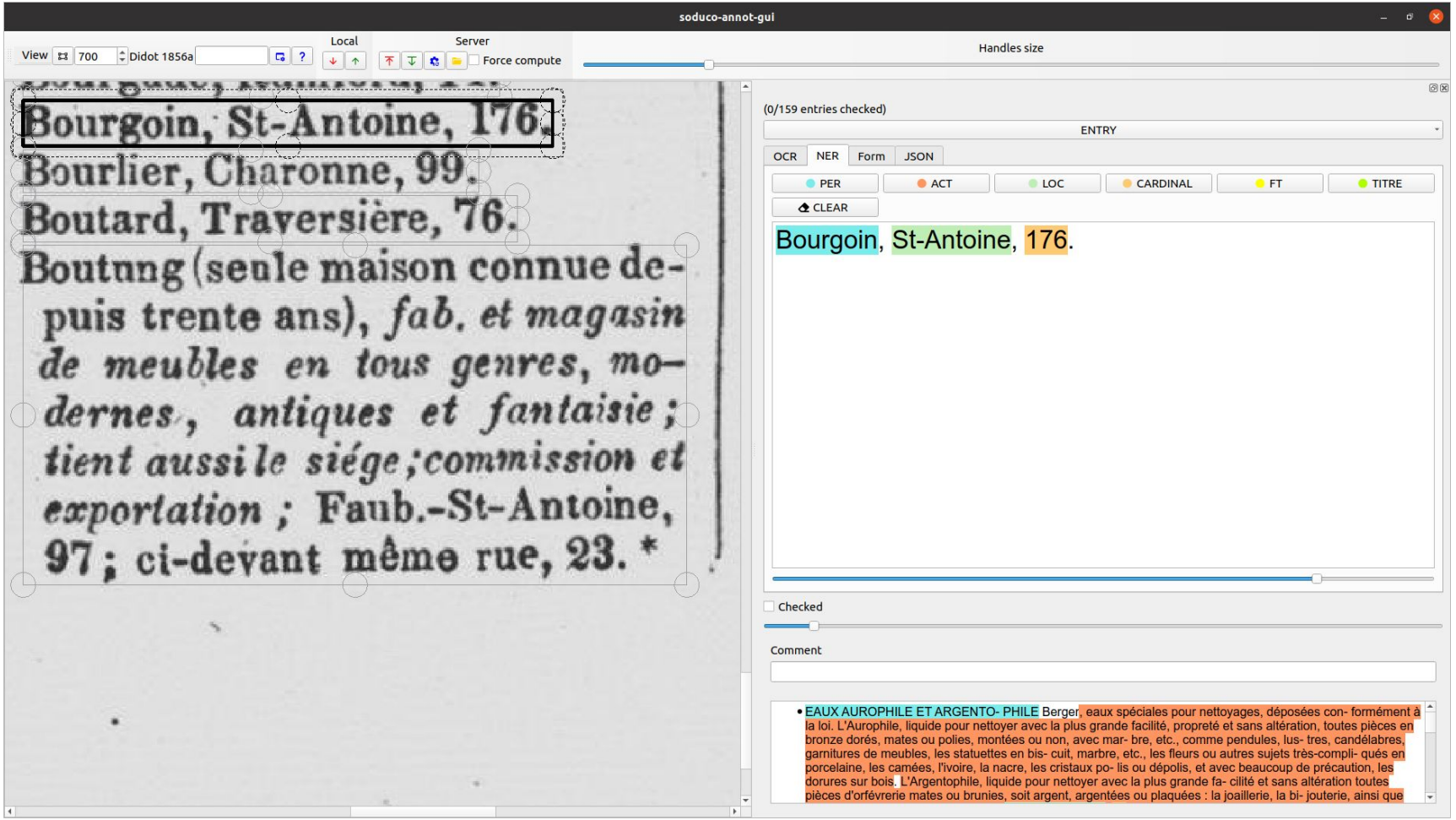

soduco-annot-gui
View
700
Didot 1856a
Local
Server
Force compute
Handles size
Boulon, Charenton, 81.
Bouquet, M. H. Exposition universelle de 1855, breveté s. g. d. g., spécialité de tables de salles à mangèr, nouveau système mécanique donnant la facilité à une seule personne de pouvoir l'ouvrir, n'importe sa grandeur, Faub.-St-Antoine, 99, passage du Bras-d'Or. *
Bourdier, fab. de divans en tous genres, fait tout ce qui concerne sa partie, Charenton, 16.
Bourdier aîné, Nve-de-Lappe, 16.
Bourgade, Rumford, 14.
(0/159 entries checked)
ENTRY
OCR
NER
Form
JSON
PER
ACT
LOC
CARDINAL
FT
TITRE
CLEAR
Boulon, Charenton, 81.
Checked
Comment
• EAUX AUROPHILE ET ARGENTO- PHILE Berger, eaux spéciales pour nettoyages, déposées con- formément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec mar- bre, etc., comme pendules, lus- tres, candélabres, garnitures de meubles, les statuettes en bis- cuit, marbre, etc., les fleurs ou autres sujets très-compli- qués en porcelaine, les camées, l'ivoire, la nacre, les cristaux po- lis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande fa- cilité et sans altération toutes pièces d'orfévrerie mates ou brunies, soit argent, argentées ou plaquées : la joaillerie, la bi- jouterie, ainsi que


soduco-annot-gui
View
700
Didot 1856a
Local
Server
Force compute
Handles size
Bouquet, M. H. Exposition uni-
verselle de 1855, breveté s. g.
d. g., spécialité de tables de salles
à mangèr, nouveau système mé-
canique donnant la facilité à une
seule personne de pouvoir l'ou-
vrir, n'importe sa grandeur,
Faub.-St-Antoine, 99, passage
du Bras-d'Or. *
Bourdier, fab. de divans en tous
genres, fait tout ce qui concerne
sa partie, Charenton, 16.
Bourdier aîné, Nve-de-Lappe, 16.
Bourgade, Rumford, 14.
Bourgoin, St-Antoine, 176.
(0/159 entries checked)
ENTRY
OCR
NER
Form
JSON
PER
ACT
LOC
CARDINAL
FT
TITRE
CLEAR
Bouquet, M. H. Exposition uni-
verselle de 1855, breveté s. g.
d. g., spécialité de tables de salles
à mangèr, nouveau système mé-
canique donnant la facilité à une
seule personne de pouvoir l'ou-
vrir, n'importe sa grandeur,
Faub.-St-Antoine, 99, passage
Checked
Comment
• EAUX AUROPHILE ET ARGENTO- PHILE Berger, eaux spéciales pour nettoyages, déposées con- formément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec mar- bre, etc., comme pendules, lus- tres, candélabres, garnitures de meubles, les statuettes en bis- cuit, marbre, etc., les fleurs ou autres sujets très-compli- qués en porcelaine, les camées, l'ivoire, la nacre, les cristaux po- lis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande fa- cilité et sans altération toutes pièces d'orfévrerie mates ou brunies, soit argent, argentées ou plaquées : la joaillerie, la bi- jouterie, ainsi que


soduco-annot-gui
View
700
Didot 1856a
Local
Server
Force compute
Handles size
Bourdier, fab. de divans en tous genres, fait tout ce qui concerne sa partie, Charenton, 16.
Bourdier aîné, Nve-de-Lappe, 16.
Bourgade, Rumford, 14.
Bourgoin, St-Antoine, 176.
Bourlier, Charonne, 99.
Boutard, Traversière, 76.
Boutung (seule maison connue depuis trente ans), fab. et magasin de meubles en tous genres, modernes, antiques et fantaisie; tient aussi le siége; commission et exportation; Faub.-St-Antoine, 97; ci-devant même rue, 23. *
(0/159 entries checked)
ENTRY
OCR
NER
Form
JSON
PER
ACT
LOC
CARDINAL
FT
TITRE
CLEAR
Bourdier, fab. de divans en tous genres, fait tout ce qui concerne sa partie, Charenton, 16.
Checked
Comment
• EAUX AUROPHILE ET ARGENTO- PHILE Berger, eaux spéciales pour nettoyages, déposées con- formément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec mar- bre, etc., comme pendules, lus- tres, candélabres, garnitures de meubles, les statuettes en bis- cuit, marbre, etc., les fleurs ou autres sujets très-compli- qués en porcelaine, les camées, l'ivoire, la nacre, les cristaux po- lis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande fa- cilité et sans altération toutes pièces d'orfévrerie mates ou brunies, soit argent, argentées ou plaquées : la joaillerie, la bi- jouterie, ainsi que


soduco-annot-gui
View
700
Didot 1856a
Local
Server
Force compute
Handles size
Bourdier aîné, Nve-de-Lappe, 16.
Bourgade, Rumford, 14.
Bourgoin, St-Antoine, 176.
Bourlier, Charonne, 99.
Boutard, Traversière, 76.
Boutung (seule maison connue depuis trente ans), fab. et magasin de meubles en tous genres, modernes, antiques et fantaisie; tient aussi le siége; commission et exportation; Faub.-St-Antoine, 97; ci-devant même rue, 23. *
(0/159 entries checked)
ENTRY
OCR
NER
Form
JSON
PER
ACT
LOC
CARDINAL
FT
TITRE
CLEAR
Bourdier aîné, Nve-de-Lappe, 16.
Checked
Comment
• EAUX AUROPHILE ET ARGENTO- PHILE Berger, eaux spéciales pour nettoyages, déposées con- formément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec mar- bre, etc., comme pendules, lus- tres, candélabres, garnitures de meubles, les statuettes en bis- cuit, marbre, etc., les fleurs ou autres sujets très-compli- qués en porcelaine, les camées, l'ivoire, la nacre, les cristaux po- lis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande fa- cilité et sans altération toutes pièces d'orfévrerie mates ou brunies, soit argent, argentées ou plaquées : la joaillerie, la bi- jouterie, ainsi que


soduco-annot-gui
View
700
Didot 1856a
Local
Server
Force compute
Handles size
Bourgade, Rumford, 14.
Bourgoin, St-Antoine, 176.
Bourlier, Charonne, 99.
Boutard, Traversière, 76.
Boutnng (seule maison connue de-
puis trente ans), fab. et magasin
de meubles en tous genres, mo-
dernes, antiques et fantaisie;
tient aussi le siége; commission et
exportation; Faub.-St-Antoine,
97; ci-devant même rue, 23. *
(0/159 entries checked)
ENTRY
OCR
NER
Form
JSON
PER
ACT
LOC
CARDINAL
FT
TITRE
CLEAR
Bourgade, Rumford, 14.
Checked
Comment
• EAUX AUROPHILE ET ARGENTO- PHILE Berger, eaux spéciales pour nettoyages, déposées con- formément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec mar- bre, etc., comme pendules, lus- tres, candélabres, garnitures de meubles, les statuettes en bis- cuit, marbre, etc., les fleurs ou autres sujets très-compli- qués en porcelaine, les camées, l'ivoire, la nacre, les cristaux po- lis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande fa- cilité et sans altération toutes pièces d'orfévrerie mates ou brunies, soit argent, argentées ou plaquées : la joaillerie, la bi- jouterie, ainsi que


soduco-annot-gui
View
700
Didot 1856a
Local
Server
Force compute
Handles size
Bourgoin, St-Antoine, 176.
Bourlier, Charonne, 99.
Boutard, Traversière, 76.
Boutnng (seule maison connue depuis trente ans), fab. et magasin de meubles en tous genres, modernes, antiques et fantaisie; tient aussi le siége; commission et exportation; Faub.-St-Antoine, 97; ci-devant même rue, 23. *
(0/159 entries checked)
ENTRY
OCR
NER
Form
JSON
PER
ACT
LOC
CARDINAL
FT
TITRE
CLEAR
Bourgoin, St-Antoine, 176.
Checked
Comment
• EAUX AUROPHILE ET ARGENTO- PHILE Berger, eaux spéciales pour nettoyages, déposées con- formément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec mar- bre, etc., comme pendules, lus- tres, candélabres, garnitures de meubles, les statuettes en bis- cuit, marbre, etc., les fleurs ou autres sujets très-compli- qués en porcelaine, les camées, l'ivoire, la nacre, les cristaux po- lis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande fa- cilité et sans altération toutes pièces d'orfévrerie mates ou brunies, soit argent, argentées ou plaquées : la joaillerie, la bi- jouterie, ainsi que

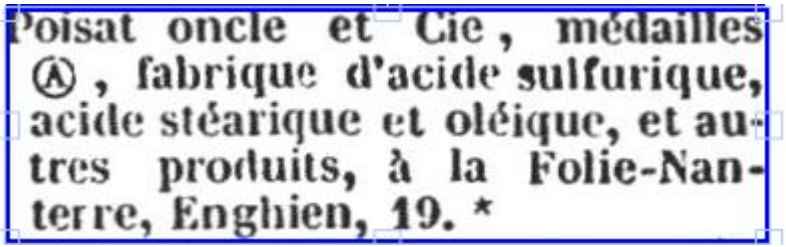
Quelques résultats qualitatifs
2. Cas limites (nouvelle interface)

Didot 1851a, vue 800


Figueran, Enghien, 7.

Figueran PER , Enghien LOC , 7 CARDINAL .

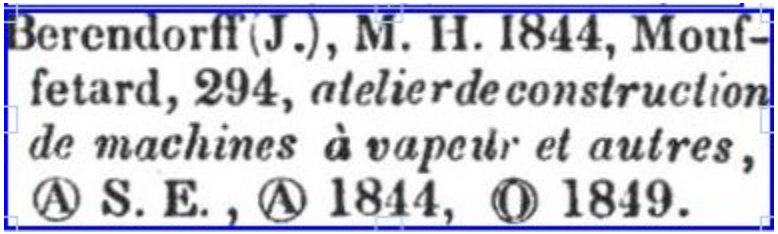
Cas typique : entrée courte avec OCR et NER fonctionnels.


Poisat oncle et Cie, médailles Ⓐ, fabrique d'acide sulfurique, acide stéarique et oléique, et autres produits, à la Folie-Nanterre, Enghien, 19. *

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

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

Didot 1851a, vue 700


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

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

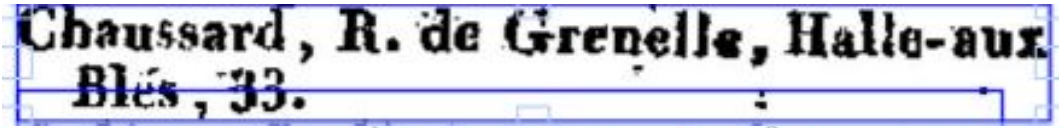
Cas typique : entrée longue avec entités principales (PER, LOC, CARD.) bien reconnues malgré l'inclusion de (“,”), et des erreurs NER sur les autres entités.

Duverneuil et La Tyna, 1806


Benard, R. des Bons-Enfans, 19.

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

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


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

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

Exemple positif, mais fusion de plusieurs niveaux de localisation

Deflandre 1829, vue 500


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

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

Exemple positif, mais fusion de plusieurs niveaux de localisation

Favre et Duchesne, 1897


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

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

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

Bottin, 1820, vue 52

des Commerçans de Paris. 49

Bourdon, *bonnetier*, r. S.-Denis, 269. (Elig.). 292

Bourdon, *md. de beurre*, r. du Faub.-S.-Honoré, 46. 318

Bourdon et co., *commiss. par eau*, q. de la Mégisserie, 38. 325

Bourdon, *épicier*, r. des Quatre-Vents, 11. 351

Bourdon (V.e), *épicière*, r. Sainte-Anne, 19. 351

Bourdon (N.) aîné et co., *miroitiers*, r. Bourg-l'Abbé, 48. 416

Bourgeois-Borlot, *plaqueur*, r. du Bac, 98. 451

Bourgeois-Dumoulin, *md. de nouveautés*, r. Bussi, 2. 410

Bourgeot, *fab. de chocolat*, r. S.-Honoré, 110. 316

Bourgeot, *papetier*, r. des Fossés-Montmartre, 31. 437

Bourget et co., *commiss. de roulage*, r. S.-Denis, 152. 323, 381

Bourget jeune, *commis. de roulage*, r. Beaurepaire, 3. 323

OCR NER

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

La détection de la structure reste une difficulté majeure.

Bilan

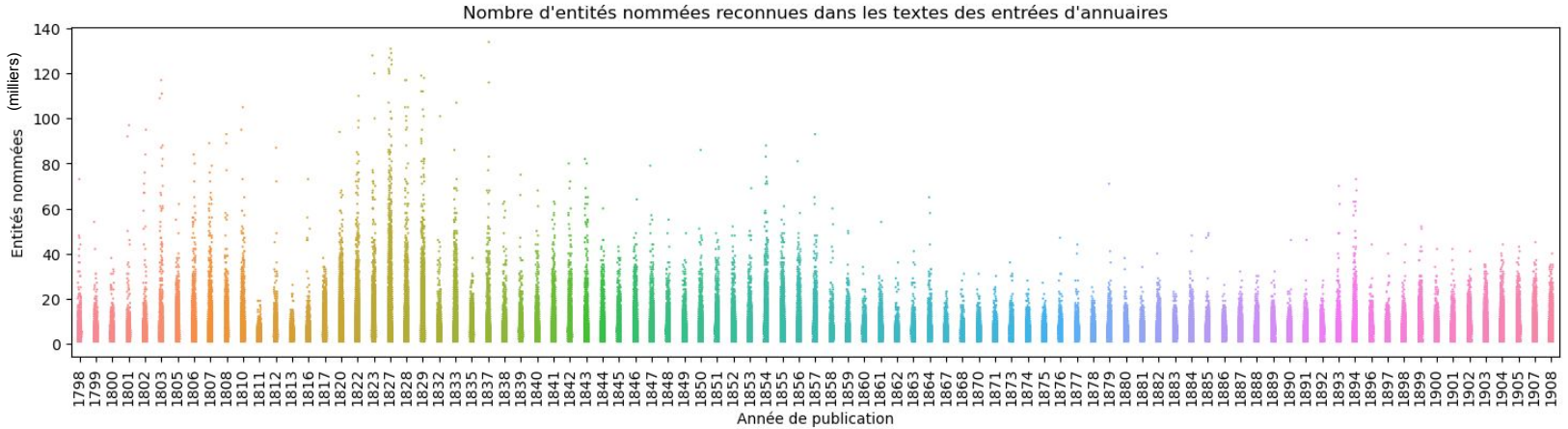
Vue d'ensemble des données extraites

Pour l'ensemble du corpus :

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

⇒ la forte redondance des entrées est riche en information

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



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

Perspectives — Projet Biblissima+

Encore une plateforme d'annotation / reconnaissance de documents imprimés ?
⇒ Non ! Vers une **collection** d'outils **ouverts**, **minimaux**, laissant les historiens **constituer leurs corpus** et y "superposer" **leurs données**.

Nos objectifs à venir :

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