A Benchmark of Named Entity Recognition Approaches in Historical Documents

Application to XIXth Century French Directories

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

(1) (2) (3)

SODUCO
ANR-18-CE38-0013

Context & motivation

Geographical Information Sciences | Urban History

Computer Sciences | Complex Networks Analysis

Social Dynamics in Urban Context: open tools, models, and data
Paris and its suburbs, 1789-1950

Pluridisciplinary research project funded by the French Research Agency, started in **2019**.

The main goal of the project is to develop methods and models to study the **evolution of the urban structure of Paris** from 1789 to 1950 in relation with social and professional practices of the city's population.

We aim to **extract large amount of fine-grained spatial and social historical data** from historical sources: large-scale maps & **trade directories**.

Trade directories of Paris, 1798-1861

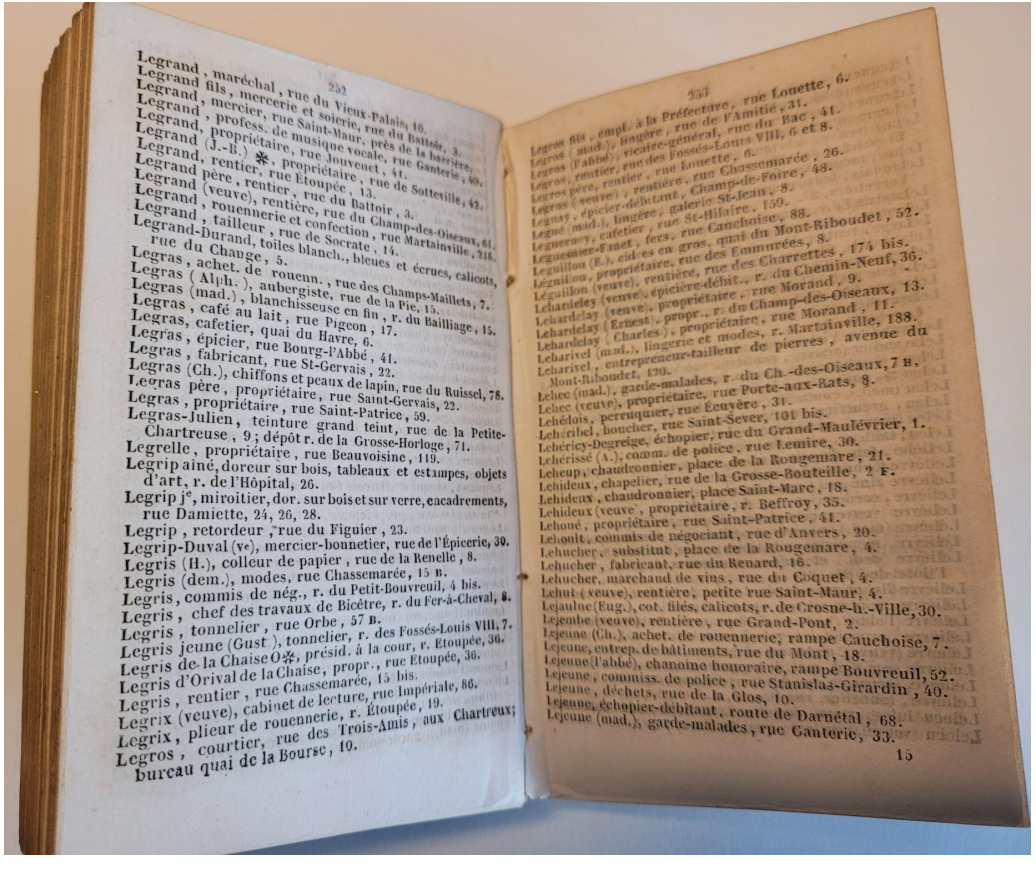
We collected **141** digitised directories.

- **100,000s** of relevant **pages**
- **1,000,000s** of raw **entries**

Directories were produced by many editors and digitised by different producers.

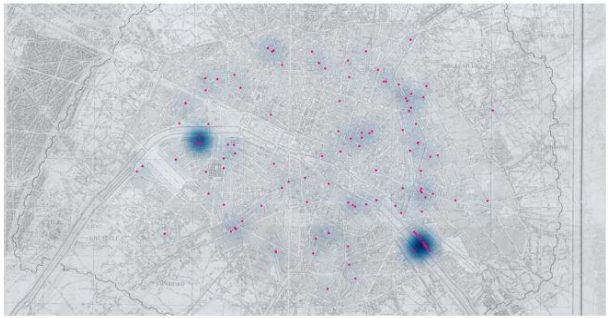
Our corpus is **heterogeneous** in content

- Length, richness and structuration of information
- Heterogenous page layouts & fonts
- Varying quality: paper thickness, digitising resolution, presence of stains, printing quirks, etc.

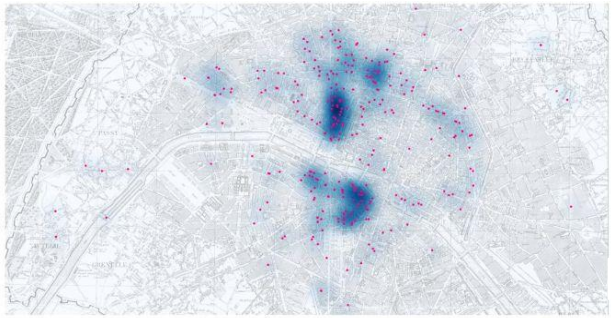


Preliminary results

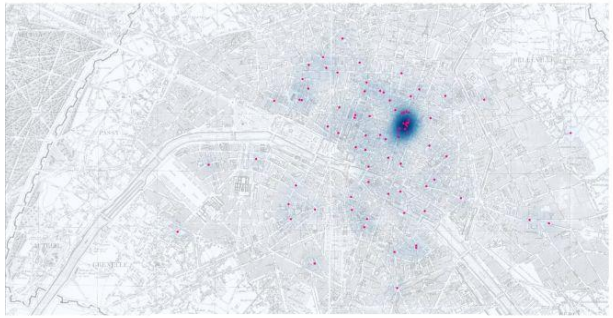
Using **addresses** extracted from **old maps** and atlases to **geocode** repositories data and analyse the evolution of the social and economic space in Paris over one century.



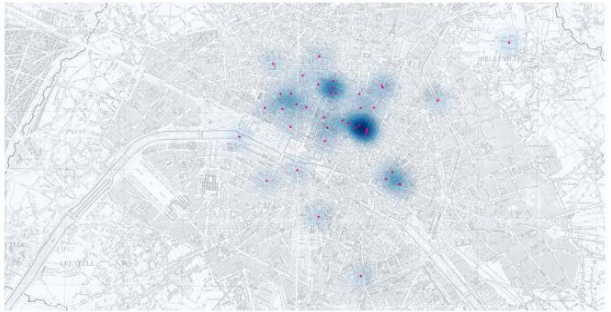
Woodworking trades



Lawyers



Umbrella manufacturers

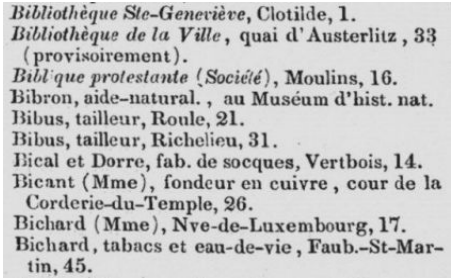


Glove makers

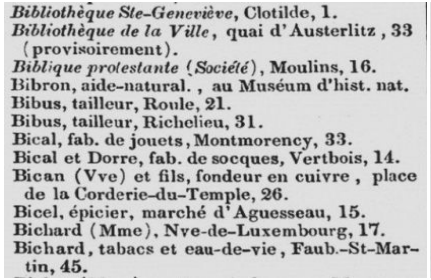
Mapping four groups of activities from the directory "Didot 1850"

DAR challenge 1: Redundancy and changes

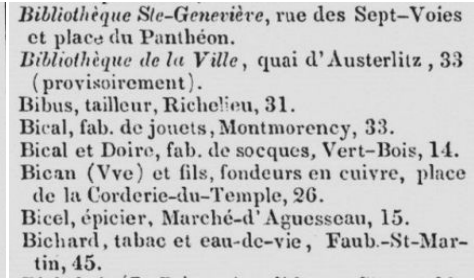
Each directory is a kind of **snapshot** of the trade actors and activities at a particular time


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.

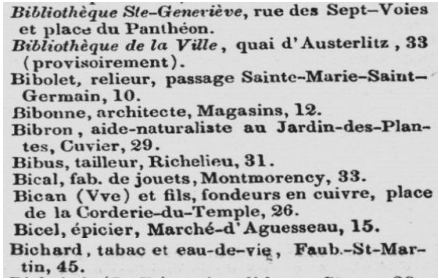
Didot 1841a - page 95


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.

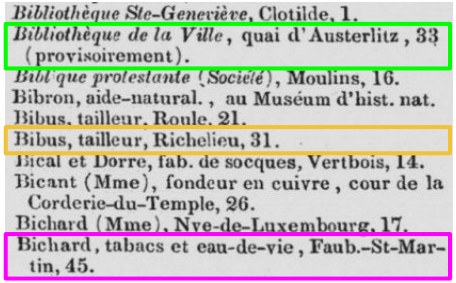
Didot 1843a - page 129


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.

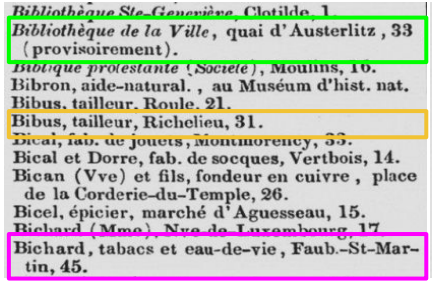
Didot 1844a - pages 125-126

DAR challenge 1: Redundancy and changes

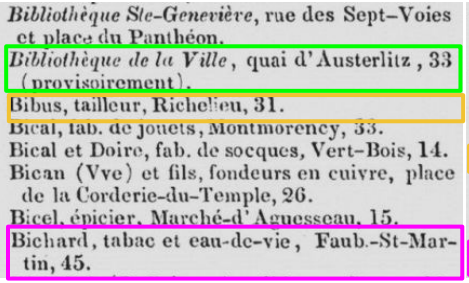
Information is **redundant** between successive directories


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.

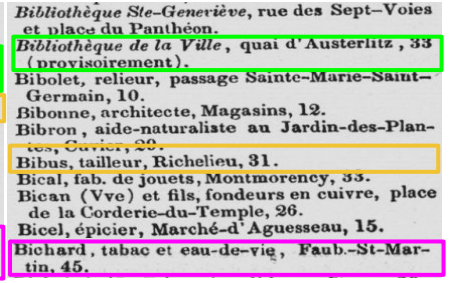
Didot 1841a - page 95


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 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.

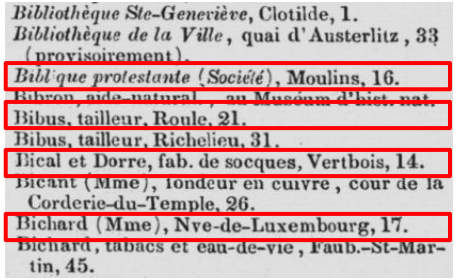
Didot 1843a - page 129


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.

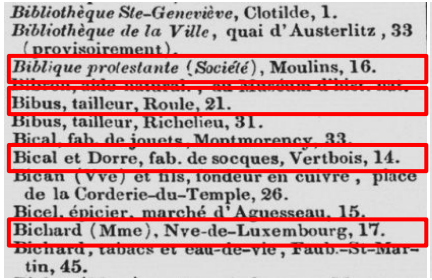
Didot 1844a - pages 125-126

DAR challenge 1: Redundancy and changes

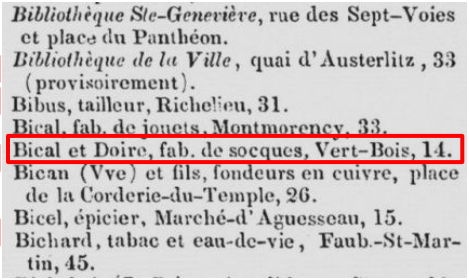
Some information **disappears** over time


Bibliothèque Ste-Geneviève, Clotilde, 1.
Bibliothèque de la Ville, quai d'Austerlitz, 33 (provisoirement).
Biblique protestante (Société), Moulins, 16.
Bibron, aide-naturalist, 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.

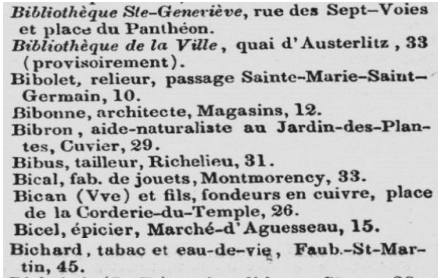
Didot 1841a - page 95


Bibliothèque Ste-Geneviève, Clotilde, 1.
Bibliothèque de la Ville, quai d'Austerlitz, 33 (provisoirement).
Biblique protestante (Société), Moulins, 16.
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.

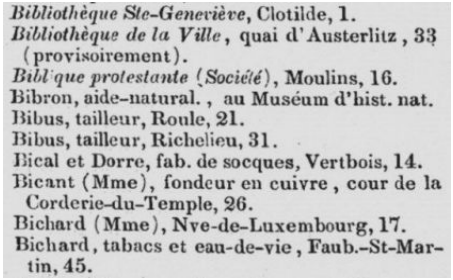
Didot 1843a - page 129


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.

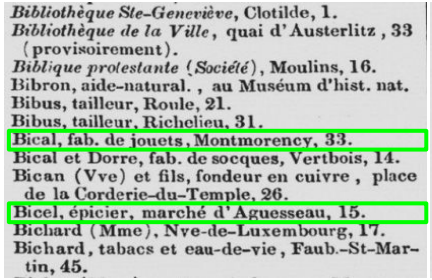
Didot 1844a - pages 125-126

DAR challenge 1: Redundancy and changes

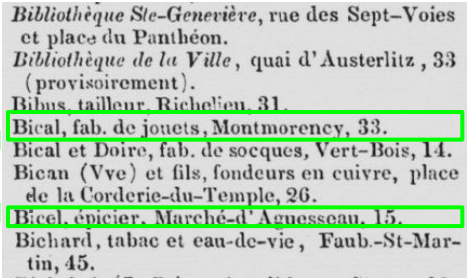
Some information **appears** over time


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.

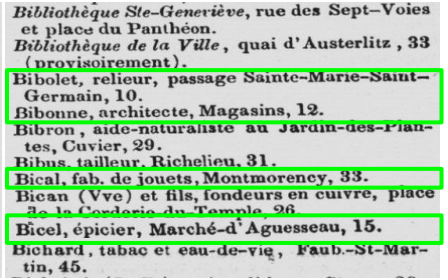
Didot 1841a - page 95


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.

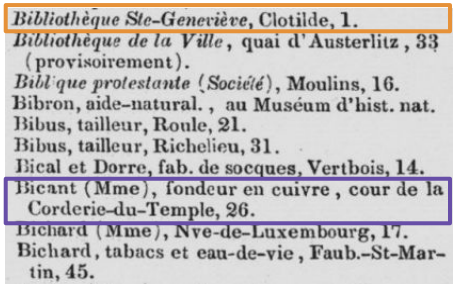
Didot 1843a - page 129


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.

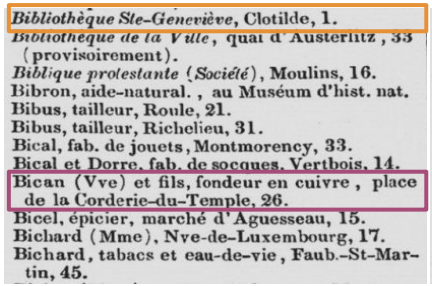
Didot 1844a - pages 125-126

DAR challenge 1: Redundancy and changes

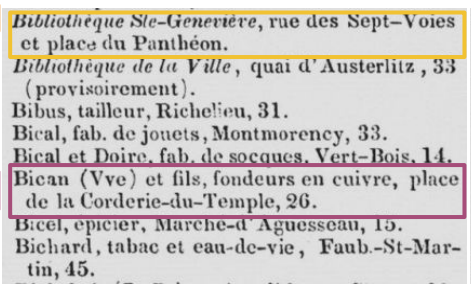
Some information **changes** over time


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.

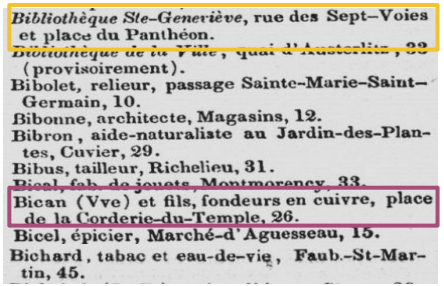
Didot 1841a - page 95


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.

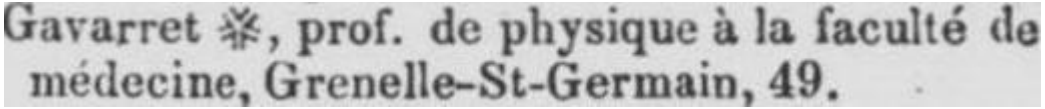
Didot 1843a - page 129


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

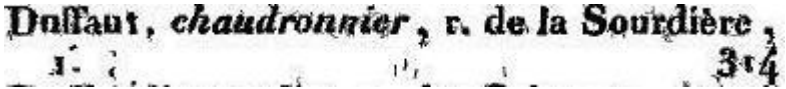
Need for deduplication.

DAR challenge 2: Unreliable transcriptions


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

= errors

Need for error detection and correction.

DAR challenge 3: Variations of entry structure

Person name | Activity | Address | Street number

Durand jeune; pour bas, Charenton, 12 ancien. 18. *

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 fourrures, 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, Fossés-St-Marcet, 12.

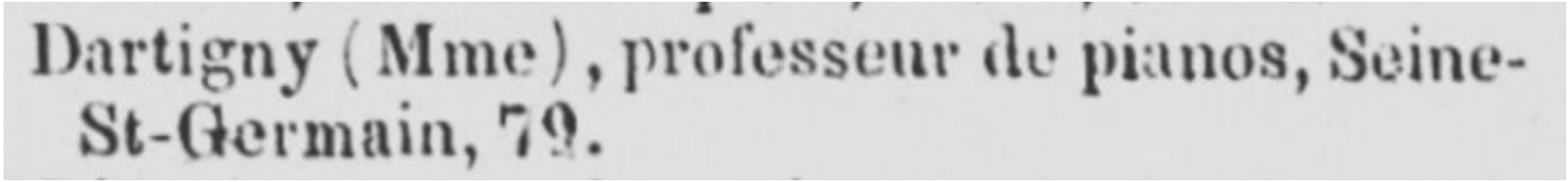
Appert fils, verres et cristaux, 21-23 Jour.

Mabire, Lourcine, 124.

Need for a robust, learned extraction system.

We need NER to extract the content of **directory entries**

Directories contain lists of **entries** somewhat similar to our “yellow pages”.


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

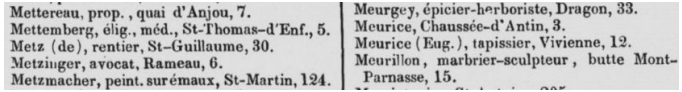
Didot 1851, page 169

↑ *typical entry in a trade directory*

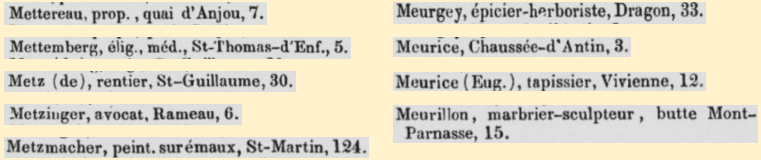
extracted entry text with named entities →

Global project pipeline

← *our focus in this paper: OCR & NER*


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.

Layout analysis


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

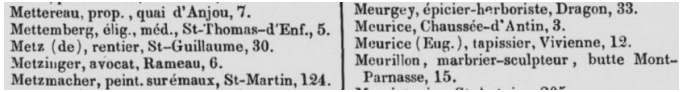
NER

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

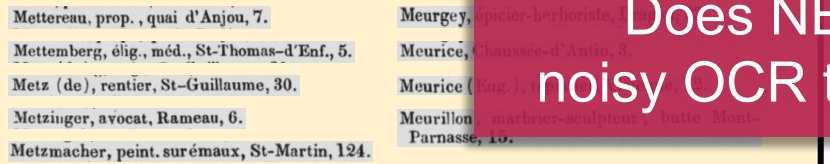
Geocoding

Global project pipeline

← *our focus in this paper: OCR & NER*


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.

Layout analysis


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.

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

NER

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

Geocoding

Does NER work on noisy OCR transcriptions?

Research problems addressed in this paper

1 OCR-related, 4 NER-related

1. What is the **performance** of **free**, **open**, **off-the-shelf OCR systems** on our directories ?
2. Which **modern**, **off-the-shelf** architectures are currently available for **NER**?
3. **How much training data** do these NER systems require?
4. Can NER system produce **meaningful results** in presence of **OCR noise**?
5. Can we **improve the tolerance** of NER systems **to OCR noise**?

OCR Benchmark:
Available options and performance

OCR systems considered

Tesseract v4 (v5 has been released since)

Pero OCR (Github code + authors 2020 model)

Kraken OCR (using a generic model for English printed text)

Evaluated as free, off-the-shelf OCR systems:

- **NO retraining**
- **NO fine-tuning**

Why? We *assumed* this would require too much *annotated* data and effort.

OCR Benchmark setup

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

A new (public, free, Zenodo-hosted) dataset DOI 10.5281/zenodo.6394464

Inputs

- 8,765 image crops (manually corrected)
- in French (Latin script with a few extra symbols: ℂ𝕄 , ✱ ...)
- 424,764 chars. in total
- extracted from 18 directories (78 pages)

Outputs — Predictions: TessV4, Pero and Kraken text

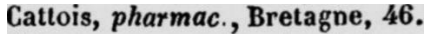
Output — Reference: Human-labelled text

Metric: CER (because WER makes little sense here) using patched UNLV-ISRI tools

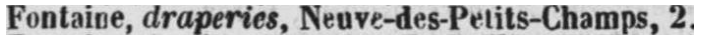
Bottin 1820

Bottin 1827

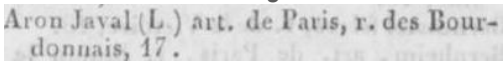
Bottin 1837



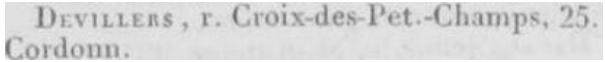
Bottin 1854



Cambon Almgene 1841

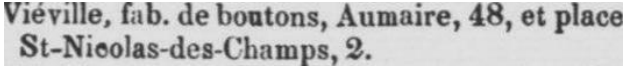


Deflandre 1828



Deflandre 1829

Didot 1851

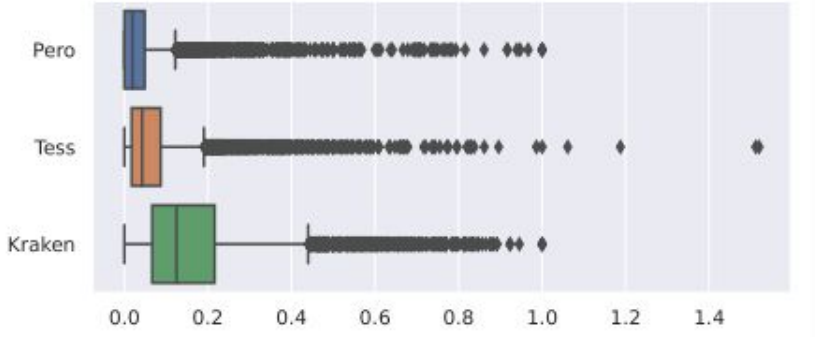


OCR Benchmark results

*Q1: How good are **free**, **off-the-shelf** OCR systems on our documents?*

A1: **Pretty good!**

	PERO OCR	Tesseract	Kraken
CER	3.78%	6.56%	15.72%



Note: we discarded Kraken in the remaining because of the low performance of the only relevant public model compatible with our data.

This is sad because Kraken has a great potential for historical documents.

NER Benchmark 1:
Available options and training data required

NER systems considered

Required because our data contains non-standard entities.

	Original training *off-the-shelf model*		**Domain adaptation** *extra training performed by us*	
SpaCy CNN:	self-supervised pre-training deep-sequoia	supervised training for NER wikiner-fr	~~pre-training~~	supervised training for NER *FTD labelled*
CamemBERT: *RoBERTa model for French*	self-supervised pre-training OSCAR	supervised training for NER wikiner-fr	~~pre-training~~	supervised training for NER *FTD labelled*
CamemBERT *pre-trained*:	self-supervised pre-training OSCAR	supervised training for NER wikiner-fr	self-supervised pre-training *FTD unlabelled*	supervised training for NER *FTD labelled*

The models are available on our HuggingFace repository and Zenodo (10.5281/zenodo.6576008).

uncorrected Pero OCR transcriptions for 845,000 raw entries (≈7000 pages)

human transcriptions and entity tags for 8765 entries (78 pages)

NER Experiment 1: setup

Input

- Clean, human-corrected text
- 8,765 entries
- 34,242 entities to detect

Outputs — NER Predictions: Spacy CNN, camemBERT (with and without pre-training)

Output — Expected: Human-labelled entities

We trained each system with **various training set sizes**:
|trainset|, |traineset|/2, |traineset|/4, ..., |traineset|/128

Goal: Assess the **amount of training data** required and **base NER performance**.

Metric: F1 score at entity level

Strict variant: a true positive (correct detection) = prediction has exactly the same start and end positions, and the same label as the expected entity

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

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

All data is available in the public FTD dataset DOI 10.5281/zenodo.6394464

NER Experiment 1: results

*Average values for **5 train/test runs***

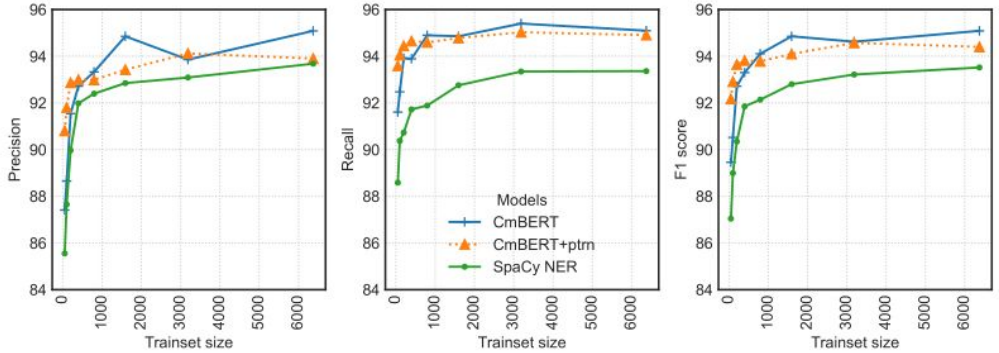
*Q2: Are there **NER options** for our problem?*

A2: Yes! And **transformer-based** models have very nice performances.

*Q3: Do they need a **lot of training data**?*

A3: No! We can get very nice results with **little data**, especially if we **leverage self-supervised pre-training**.

	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



NER Benchmark 2: Robustness to OCR noise

NER systems considered

*We **focused on camemBERT** for this experiment.*

	Original training *off-the-shelf model*		**Extra training** *performed by us*	
~~**SpaCy CNN:**~~	~~self-supervised pre-training deep-sequoia~~	~~supervised training for NER wikiner-fr~~	~~pre-training~~	~~supervised training for NER *FTD labelled*~~
CamemBERT: *RoBERTa model for French*	self-supervised pre-training OSCAR	supervised training for NER wikiner-fr	~~pre-training~~	supervised training for NER *FTD labelled*
CamemBERT *pre-trained*:	self-supervised pre-training OSCAR	supervised training for NER wikiner-fr	self-supervised pre-training *FTD unlabelled*	supervised training for NER *FTD labelled*

The models are available on our HuggingFace repository.

NER Experiment 2: setup

Input
- Clean, human-corrected text
- or **noisy OCR** (TessV4 and Pero)

Outputs — Predictions
- NER predictions from camemBERT
  Without and with pre-training

Output — Reference
- human-labeled entities
- or **projected human annotations on noisy OCR text**

Metric: F1 score
(strict variant, as in NER XP 1)

Also available in the public FTD dataset DOI 10.5281/zenodo.6394464

***camemBERT pretrain/train/test variations** (× 12)*

Pretrain set (× 2)	Train set (× 2)	Test set (× 3)
• *None* • PERO (raw)	• Reference • PERO (proj. annot.)	• Reference • PERO • Tesseract

Example of reference NER outputs

Human-labelled:
Dulay ***PER***, chandronnier ***ACT***, r. du Pont- aux Choux ***LOC***, 15 ***CARDINAL***, 314.

Entities **projected** on **Tesseract's** output:
Dulay ***PER***, chandronnier ***ACT***, +. du Pont-anx-Cars ***LOC*** Ge 7 ***CARDINAL*** Fe ÊR one

NER Experiment 2: results

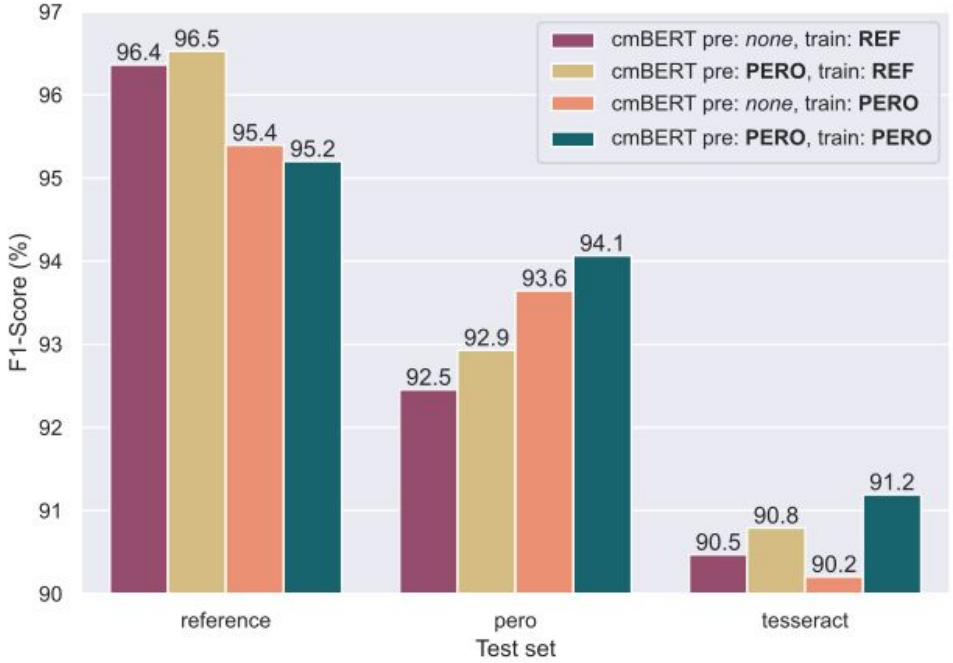
*Average values for **5 train/test runs***

*Q4: Can NER work on **noisy OCR transcriptions**?*

A4: **Yes!** (but worse than with clean data)

*Q5: Can we make NER **more robust** to noise?*

A5: **Yes! Pretrain** (self-supervised) AND **train** on (supervised using projected reference) on the same **noisy OCR data**
⇒ don't train on clean text if you target noisy OCR!



Take-home messages

Benchmark of NER approaches on noisy OCR predictions

OCR

1. Modern free, off-the-shelf OCR are **good**.

NER

2. **NER** architectures based on **transformers** are very efficient…
3. …and **cheap to train**.

4. They still can isolate entities correctly even in presence of **noisy OCR predictions**…
5. …and there performance can be improved by leveraging **self-supervised pre-training** and **supervised training on noisy data** (requires to project human annotations over noisy text).

But there's more:

The **FTD dataset**, including all inputs and outputs presented here is available freely at https://zenodo.org/record/6394464

- Suitable for **OCR evaluation**
- Suitable for **NER fine-tuning** *self-supervised and supervised*
- Suitable for **NER evaluation**

All our code, including **NER training**, is freely available at https://github.com/soduco/paper-ner-bench-das22

Please reuse, fork, improve them!