Efficient NLP for Clinical Data Warehouses

EDS-NLP and some usecases

03/07/2025

AI4HealthSummer School 2025 @ Owkin

Perceval Wajsbürt, PhD
Data Science Team
Innovation et Données, DSN, AP-HP

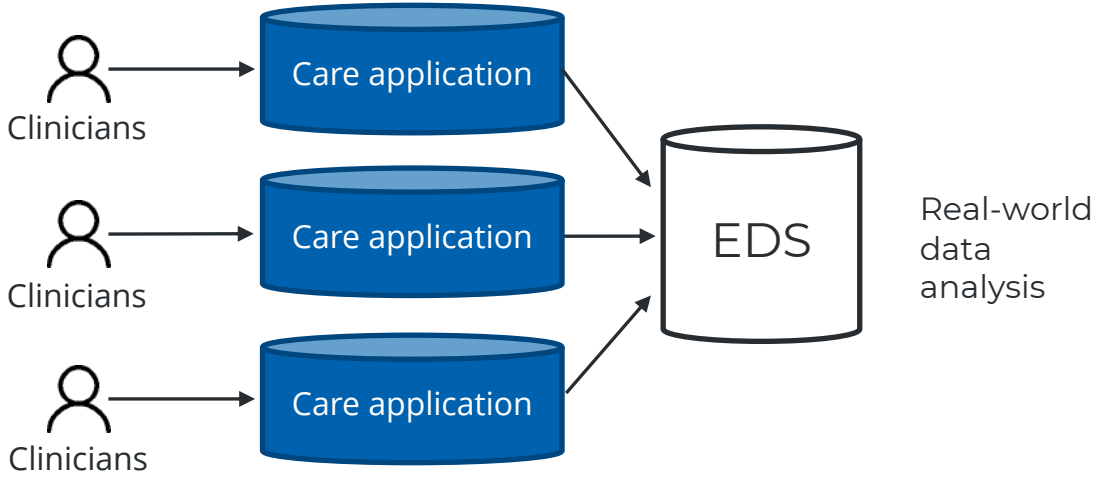
ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Summary

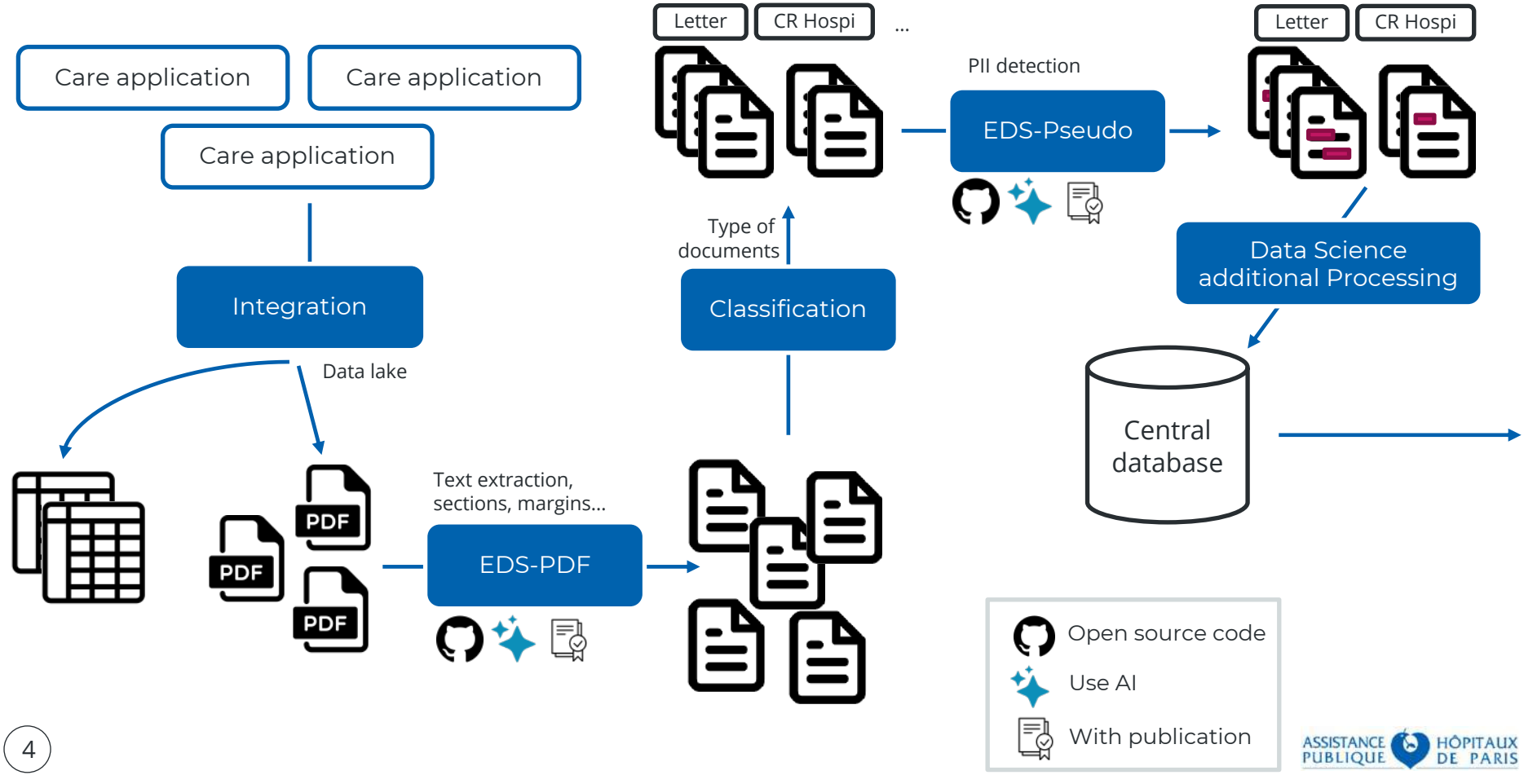
Quick context	• Processing of documents à AP-HP's CDW
EDS-NLP	• Need for a common tool • Equation to solve • Overview and collaborative approach • Some examples • Acceleration of inference
Focus on pseudonymization	• Context and objectives • Hybrid solution • Document annotation • Replacements • Results and public model

The EDS of AP-HP: a database

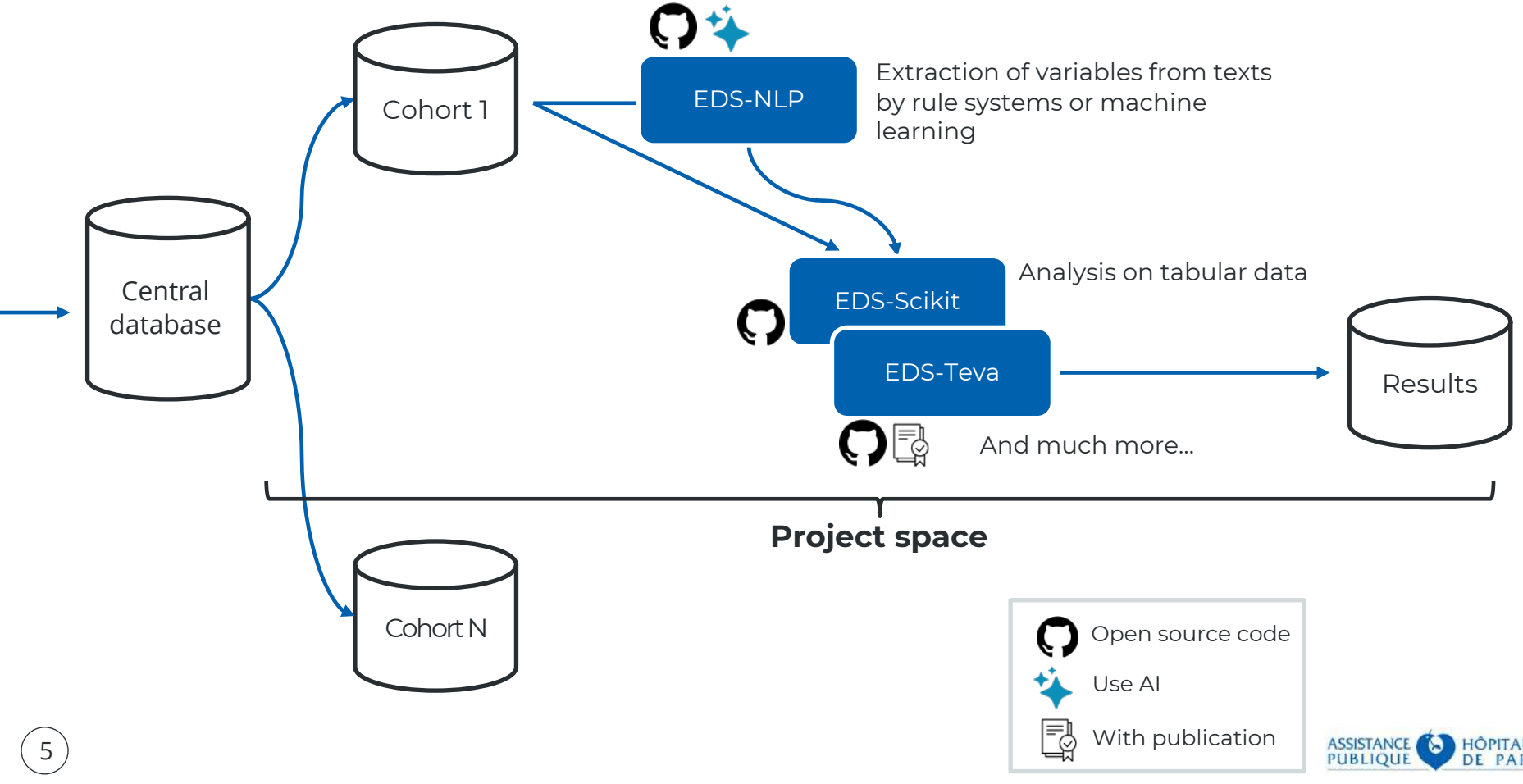
The Clinical Data Warehouse of AP-HP (**EDS**) is a database that contains real-world data collected by the software of AP-HP.
Objectives: **research**, **innovation,** and **management**
The research information system is directly connected to the clinical information system.



Processing of CDW documents : upstream by AP-HP's IT Teams



Processing of CDW documents: downstream by the projects



EDS-NLP: Need for consolidation of NLP projects at EDS

Findings

- NLP is increasingly used by data scientists, biostatisticians, clinician-researchers.
- Numerous cross-cutting algorithms: diabetes, prognostic score, prescription, smoking...
- High-performing rule-based approaches but time-consuming for clinicians.
- Regular evolution of algorithms to account for special cases and developments of the information system.
- The reproducibility of studies is an important component of their quality.
- Published algorithms are mainly in English.

Objective

Create a reliable, versioned scientific library for French clinical NLP that is co-developed by a community of diverse stakeholders from research, innovation, and management.

EDS-NLP: Equation to solve

Use cases

- Specific search for terms ("TS")
- Prescription extraction, diagnosis
- Extraction of dates, quantities, scores
- Negation, hypothesis, kinship
- Detection of relationships
- Normalization: conversion of quantities, date parsing, terminology matching

+

Nature of clinical texts

- PDF extraction artifacts "======="
- Great diversity of formulations:
  - *The patient is not diabetic.*
  - *15 pack-years ... / He smokes*
  - *Patient is 1.50m / Height: 150cm*
- Special structures: sections, lists

+

Ecosystem constraints

- Formats: Spark / Parquet / Pandas
- Large volumes of data
- Computing infrastructure
  - *CPU/GPU cluster*
  - *"Fragile network"*
- Share models, without risk for patients

+

Wide variety of models

- Rule-based models
- ML Models
- Hybrid models
- Modularity
- Allow custom ML models
- But also their sharing*

= EDS-NLP

EDS-NLP: Overview

Pipeline system

- API inspired & compatible w/ spaCy
- Deep learning with Pytorch
- Modular component pipeline system

```python
import edsnlp

nlp = edsnlp.blank('eds')
nlp.add_pipe('eds.normalizer')
nlp.add_pipe('eds.covid')
nlp.add_pipe('eds.negation')
```

More than 50 bricks based on rules

- Preprocessing: sentence segmentation, sections, text cleaning
- generic extractors: regex, lexicons
- specific extractors: comorbidities, medications, ICD-10, TNM, Adicap, ...
- date extractors, measurements
- qualification: negation, kinship, hypothesis

Trainable deep-learning bricks

- Embeddings: Transformer, CNN, Pooler
- Nested NER
- Multi-label entity classification
- Normalization of named entities
- And more...

EDS-NLP: A collaborative environment

- Versioned project, documented, tested online, licensed under BSD 3-Clause:

- Discussions, bug reports, and improvements **through GitHub**
- Citation of scientific publications in the documentation

Authors and citation

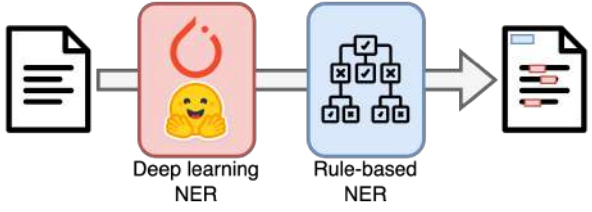
The TNM score is based on the development of S. Priou, B. Rance and E. Kempf (Kempf et al., 2022).

1. Kempf E., Priou S., Lamé G., Daniel C., Bellamine A., Sommacale D., Belkacemi y., Bey R., Galula G., Taright N., Tannier X., Rance B., Flicoteaux R., Hemery F., Audureau E., Chatellier G. and Tournigand C., 2022. Impact of two waves of Sars-Cov2 outbreak on the number, clinical presentation, care trajectories and survival of patients newly referred for a colorectal cancer: A French multicentric cohort study from a large group of University hospitals. *{International Journal of Cancer}. 150*, pp.1609-1618. 10.1002/ijc.33928

EDS-NLP: Some examples of projects

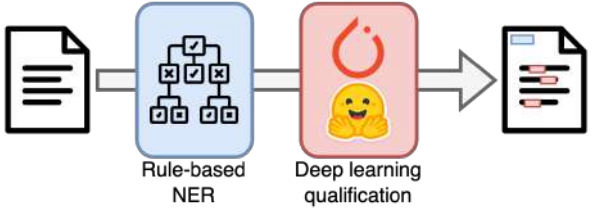
EDS-Pseudo

Detection of identifying entities



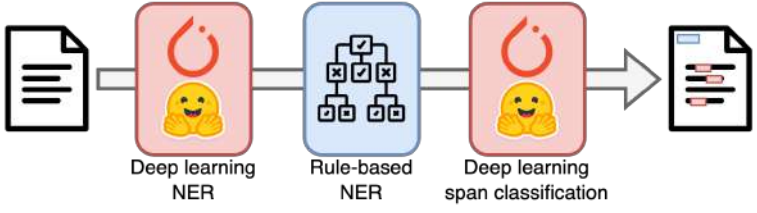
Comorbidities detection

Detection of Charlson comorbidities



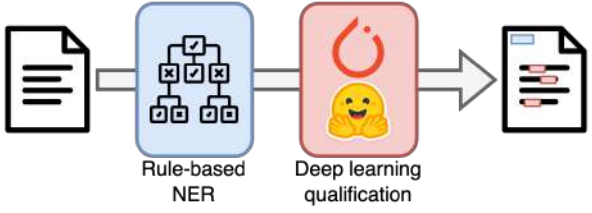
EDS-Medic

Extraction of medication prescriptions

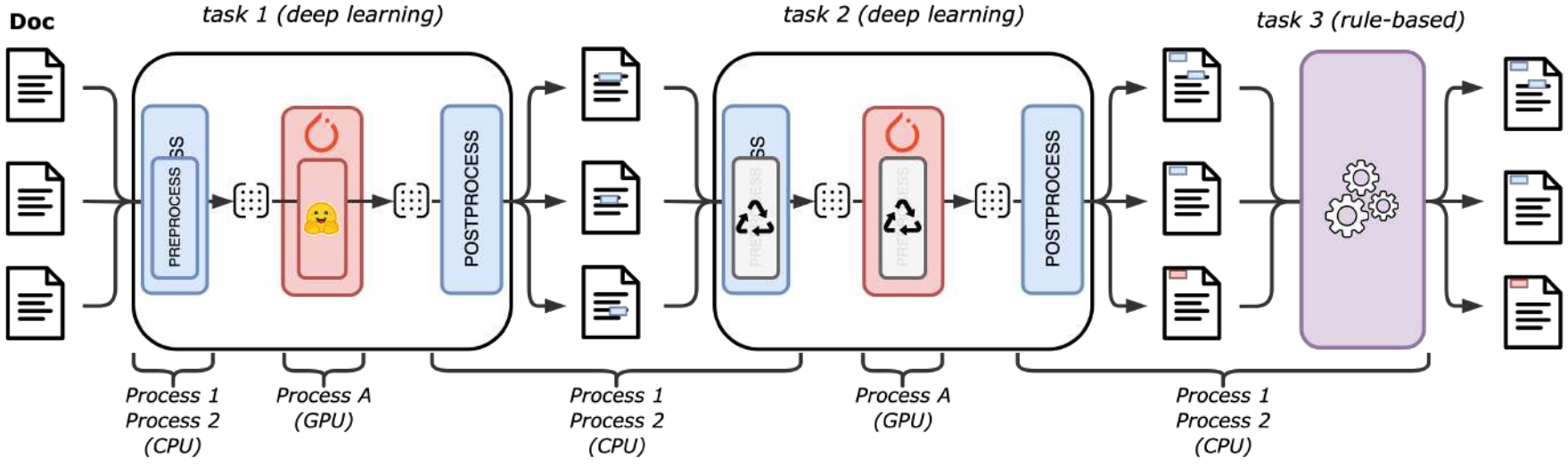


Suicide attempt detection

Detection of suicide attempts



EDS-NLP: Multi-GPU inference parallelization



- Objectives: GPUs at 100% utilization and vectorization of calculations
- Minimization of data exchanges
- We parallelize the pre-/post-processing and rule-based components.

EDS-NLP: Results of the parallelization of EDS-Medic

# GPU	# CPU	Doc/s	Speedup	eq.gCO2/100k
1	0	80	1,0	74
1	1	102	1,3	
1	2	195	2,4	43
1	3	277	3,4	
1	4	341	4,2	37
1	6	373	**4,6**	33
2	4	390	4,9	
2	8	682	8,5	
4	16	1184	14,7	
4	24	1359	16,9	

⇦ Equivalent maximum spaCy stable

- Tests on 1 to 4 A100 GPUs
- EDS-NLP allows for better use of resources.
- Increased inference speed
- And a reduction in energy costs

Pseudonymisation

Context

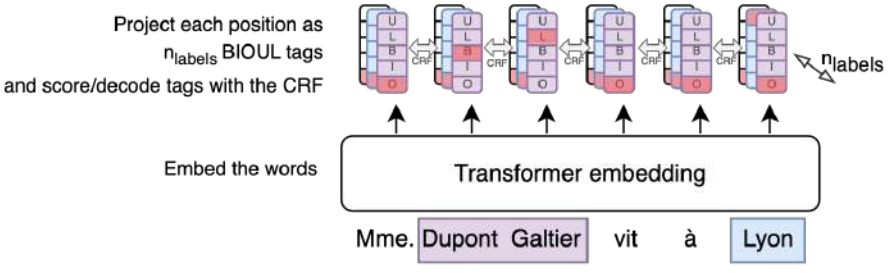
- The texts of the EDS contain numerous **identifying information.**
- The CNIL asks the EDS to **keep up with the SOTA** to remove as much of this information as possible.
- It is known that **ML** allows for better pseudonymization performance.

Challenges

- Develop a state-of-the-art pseudonymization algorithm for clinical trial reports.
- **Production setup** for the daily integration of clinical texts into the EDS.
- **Validate** results ⇨ ensure transparency regarding **residual risks**
- **Open-source** the developments made to open the project to collaborations.

Pseudonymisation: Solution

1. **Deep-learning**: Transformer + CRF



2. **Rules**
   - Statics (regex ++)
   - Dynamics: research based on structured data
3. **Fusion of ML results** + Rules

4. Implementation

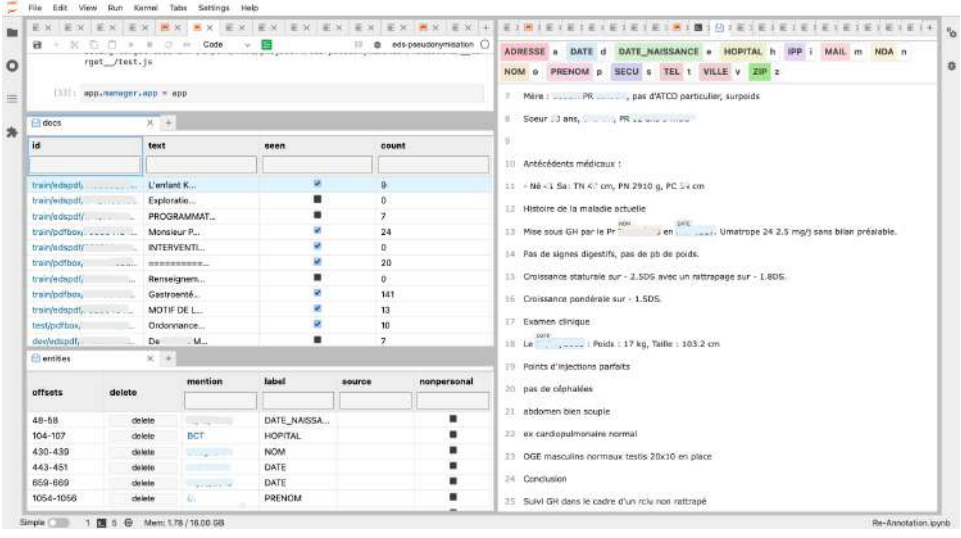

```python
import edsnlp

nlp = edsnlp.blank('eds')
nlp.add_pipe(
  'eds.ner_crf', # to be trained
  config={
    "embedding": {
    "@factory": "eds.transformer",
    ...}
})
nlp.add_pipe('eds.dates')
nlp.add_pipe('eds_pseudo.adresses')
nlp.add_pipe('eds_pseudo.misc_rules')
nlp.add_pipe('eds_pseudo.context')
```

Pseudonymisation: Annotation

Two-phases campaign

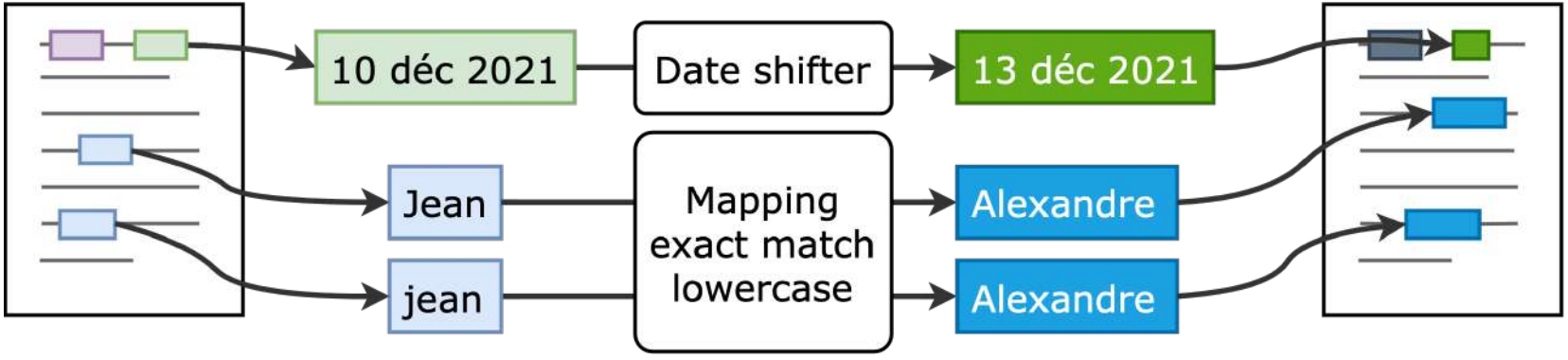
- First phase in Dec 21 - Jan 22
- Correction in Dec 22 - Jan 23

	ENTS	DOCS
train/edspdf	27135	3025
train/pdfbox	16071	348
dev/edspdf	1615	200
dev/pdfbox	967	22
test/edspdf	**3491**	348
test/pdfbox	**16793**	348



2nd phase of annotation (metanno)

Pseudonymisation: Replacement



Instead of masking, we replace, which helps to hide errors among the correctly replaced data.

- Replacement using exact matches in lowercase
- ⚠ Be sure to adjust the indices if entity detection is done beforehand.

Pseudonymisation: Results

Metrics

- Word Level (not an exact match)
- Redact: % of words redacted
- Full: % of docs with 100% PII redacted

Results

- 99.0 F1 score
- 99.4% of PII words redacted
- 86.2% of docs fully redacted
- VISIT ID difficult because of diverse formats and truth in structured data

Label	P	R	F1	Redact	Full
ADDRESS	99.0	98.4	98.7	98.5	98.4
BIRTHDATE	98.2	98.2	98.2	99.8	99.7
CITY	98.0	98.8	98.4	98.8	98.2
DATE	99.7	99.3	99.5	99.6	95.4
EMAIL	98.9	99.9	99.4	99.9	99.9
FIRSTNAME	98.8	98.4	98.6	99.4	97.4
LASTNAME	98.6	98.6	98.6	99.6	97.2
NSS	88.0	98.9	93.1	100.	100.
PATIENT ID	99.0	94.0	96.4	98.2	99.1
PHONE	99.6	99.7	99.7	99.7	99.0
VISIT ID	91.5	89.4	90.4	90.4	98.3
ZIP	99.9	99.9	99.9	99.9	99.9
ALL	99.0	98.9	**99.0**	99.4	86.2

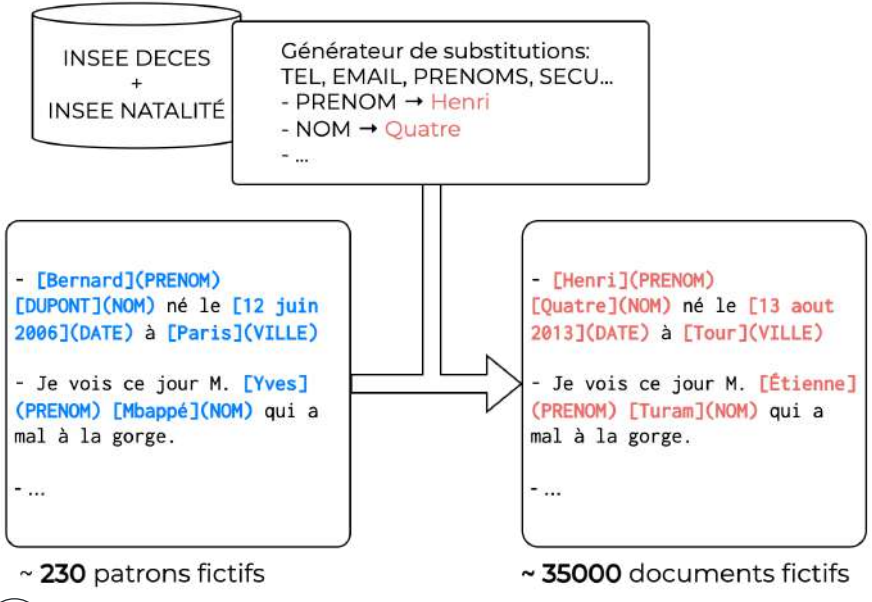
NEW !

eudonymisation: Public model

Model: hf.co/AP-HP/eds-pseudo-public

Demo: eds-pseudo-public.streamlit.app

Generation of fictional examples:



Label	P	R	F1	Redact	Full
ADDRESS	98.2	96.9	97.6	97.6	96.7
BIRTHDATE	97.5	96.9	97.2	99.3	99.4
CITY	96.7	93.8	95.2	95.1	91.1
DATE	99.0	98.4	98.7	98.8	85.9
EMAIL	96.1	99.8	97.9	99.8	99.7
FIRSTNAME	93.5	96.6	95.0	99.0	93.2
LASTNAME	94.4	95.3	94.8	98.2	89.5
NSS	88.3	100	93.8	100	100
PATIENT ID	91.9	90.8	91.3	98.5	99.3
PHONE	97.5	99.9	98.7	99.9	99.6
VISIT ID	92.1	83.5	87.6	87.4	97.2
ZIP	96.8	100	98.3	100	100
ALL	97.0	97.8	**97.4**	98.8	63.1

Do you have any questions?