SUMAC @ ACM 2025, DUBLIN - IRELAND

Benchmarking OCR Tools for Historical Postcards: A Dataset and Evaluation

Matthieu ***PELINGRE****, UL (LORIA) - Nancy (FR)*

Salvatore ***TABBONE****, UL (LORIA) - Nancy (FR)*

SUMAC 2025

ACM multimedia

Dublin, Ireland 27-31.10.2025

Loria

UNIVERSITÉ DE LORRAINE

INTRODUCTION

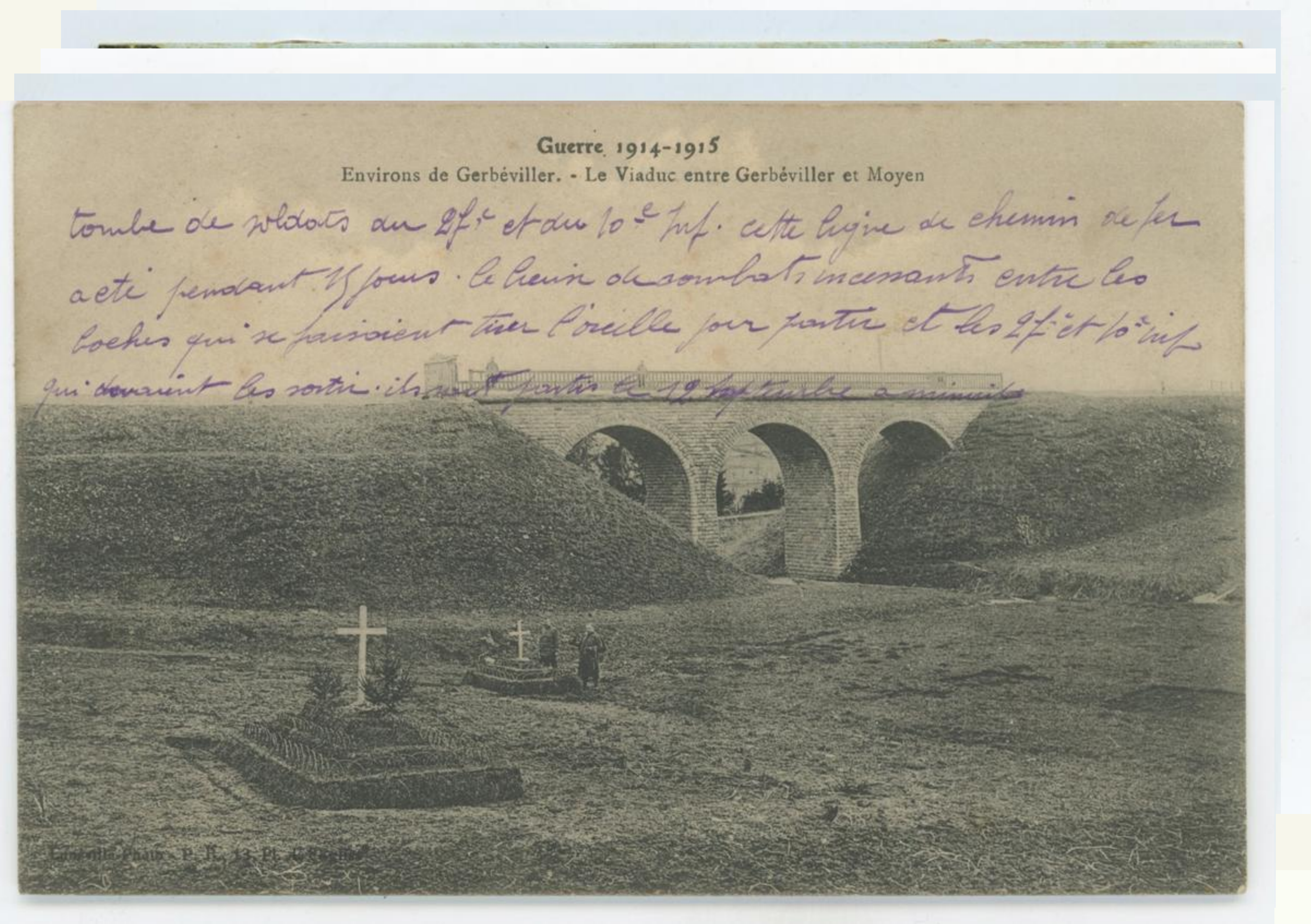
A LINK TO THE PAST

IMAGE'EST

4,293 postcards

1890s → 1930s

Visualy rich



Guerre 1914-1915

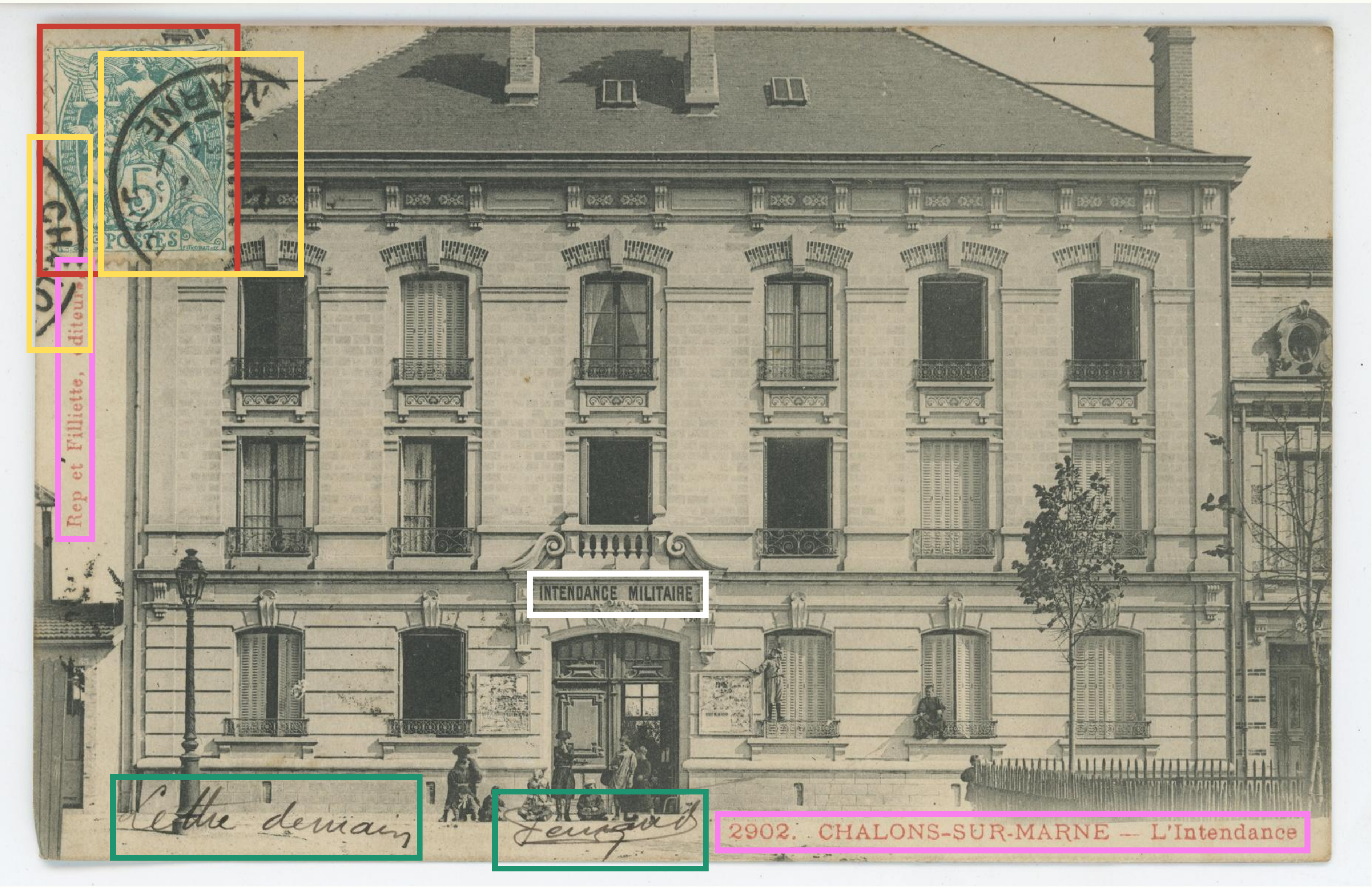
Environs de Gerbéviller. - Le Viaduc entre Gerbéviller et Moyen

tombe de soldats du 27e et du 10e Inf. cette ligne de chemin de fer
a eté pendant 15 jours le lieu de combats incessants entre les
boches qui se faisaient tirer l'oreille pour partir et les 27e et 10e inf
qui devaient les sortir. ils sont partis le 12 septembre a minuit

Lunéville Phot. P. H. 13 Pl. [illegible]

PROBLEM AND OBJECTIVES

Few OCR-Ready Historical Postcard Data Sets


INTENDANCE MILITAIRE
2902. CHALONS-SUR-MARNE — L'Intendance
Rep et Filliette, éditeurs
Lettre demain
Bernard
POSTES
5c

Printed 1: 2902. CHALONS-SUR-MARNE – L'Intendance

Printed 2: Rep et Filliette, éditeurs,

Handwritten 1: Lettre demain

Handwritten 2: Bernard

Scene text: Intendance Militaire

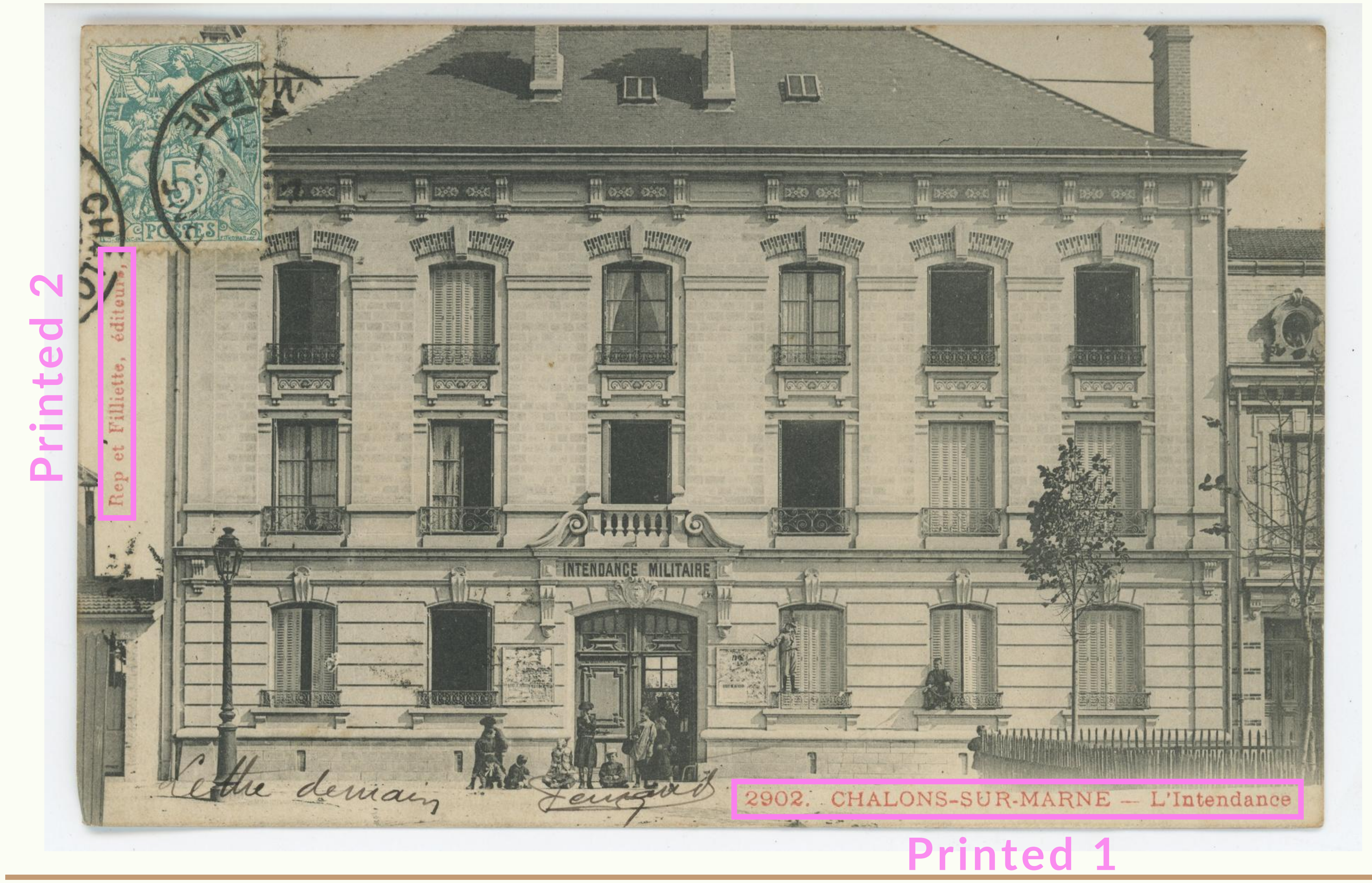
Postage Stamp: "Blanc" type – 5c

Date Stamp: Châlons-sur-Marne

➔Complex: ≠ orientations and fonts, overprinting …

PROBLEM AND OBJECTIVES

Focus on Printed Text


Printed 2
Rep et Filliette, éditeurs,
INTENDANCE MILITAIRE
2902. CHALONS-SUR-MARNE — L'Intendance
Printed 1

Printed 1: 2902. CHALONS-SUR-MARNE – L'Intendance

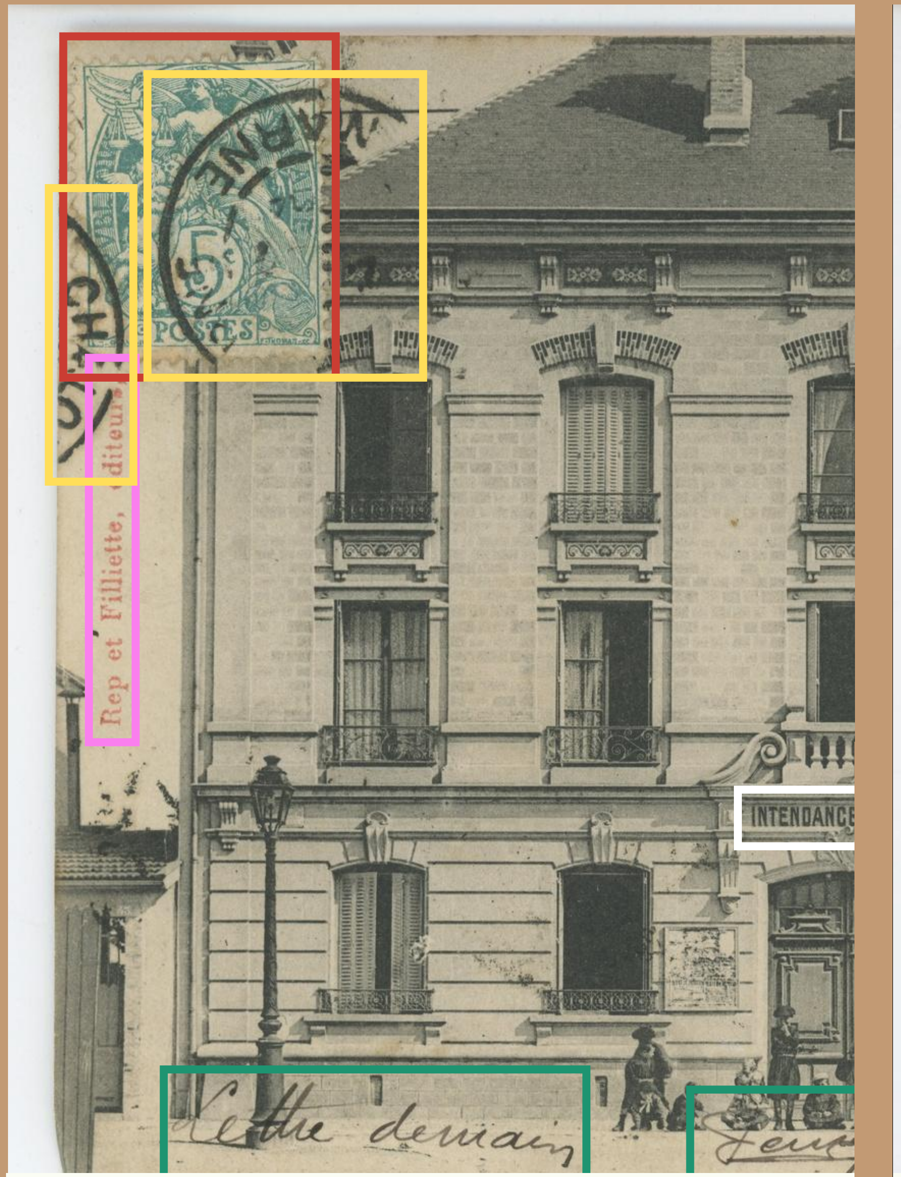
> **Photo location:** Châlons-sur-Marne
>
> **Current zipcode:** 51 000
>
> **Keyword:** Intendance (Supply Corps)

Printed 2: Rep et Filliette, éditeurs,

> **Editor:** Rep et Filliette
>
> **Edit. city:** Château-Thierry
>
> **Keyword:** éditeur (editor)

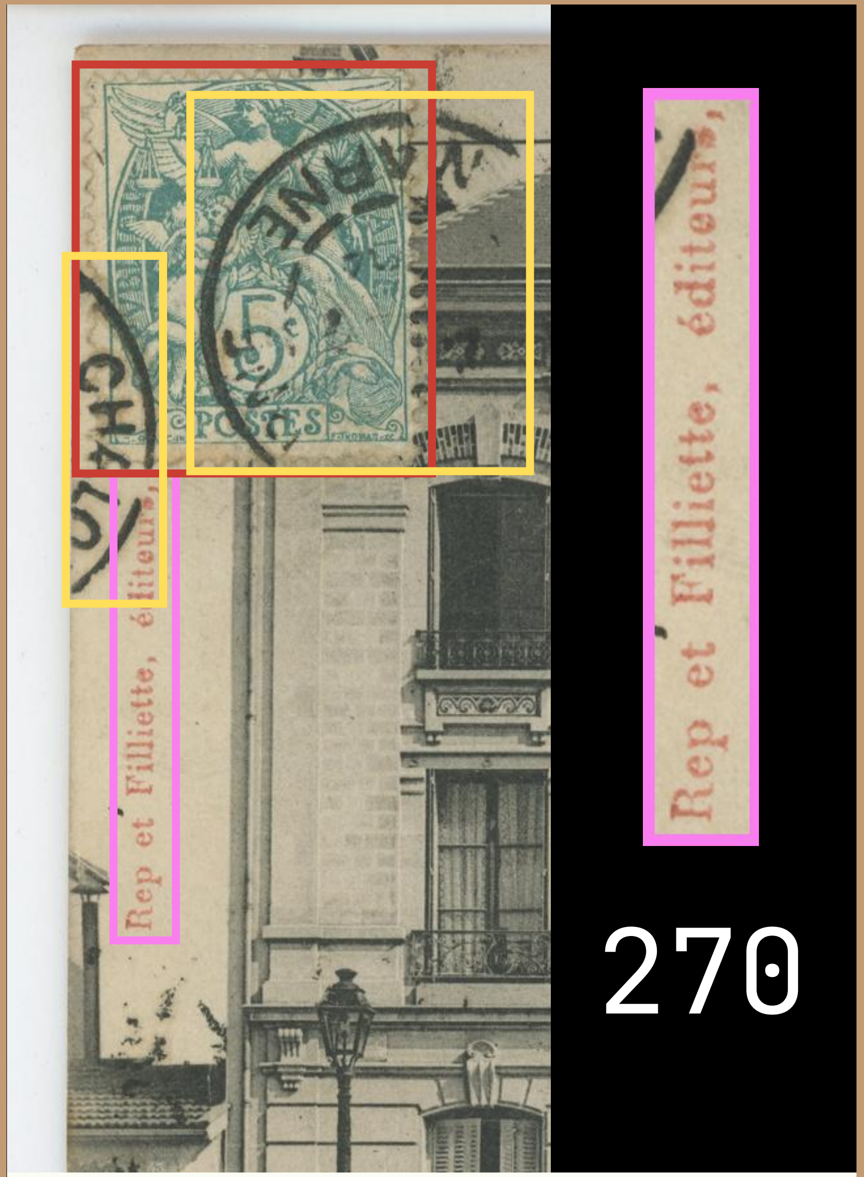
➔ Handwritten and scene texts in future updates

OCR PIPELINE



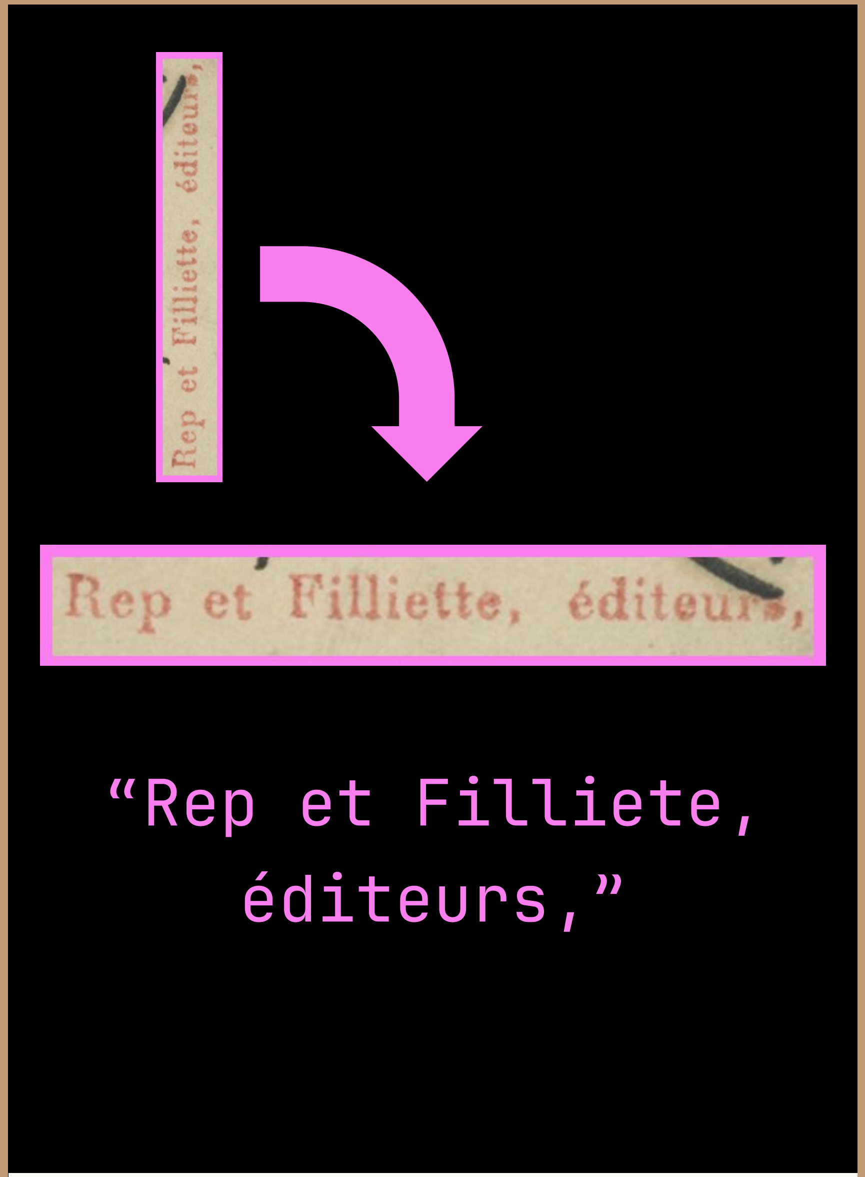
DETECTION

YOLOv12



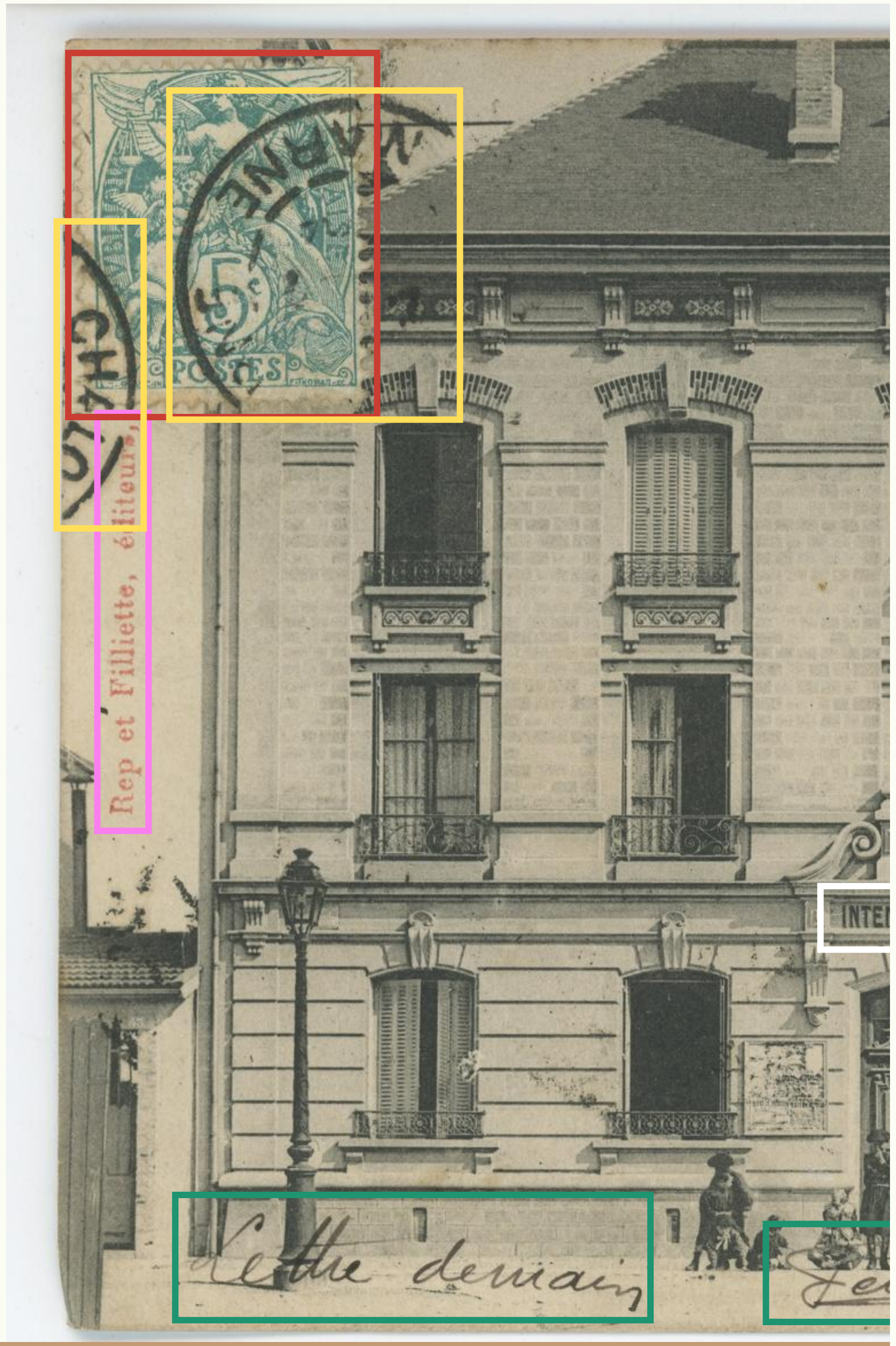
ORIENTATION

Tesseract OCR



TRANSCRITPION

EasyOCR



STEP 1: DETECTION

TEXT & POSTAL MARKINGS DETECTION

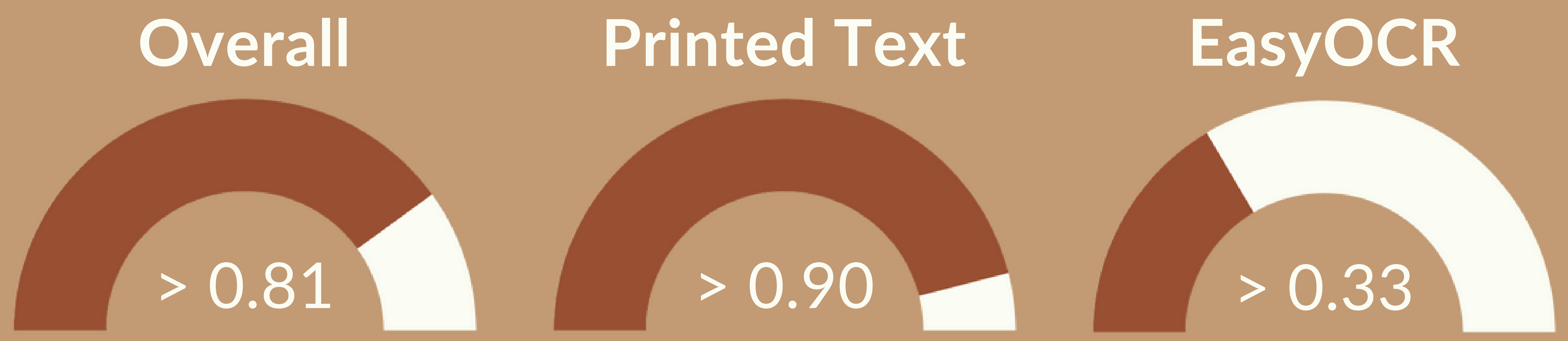
Fine-tuned YOLOv12 (1,375 postcards)

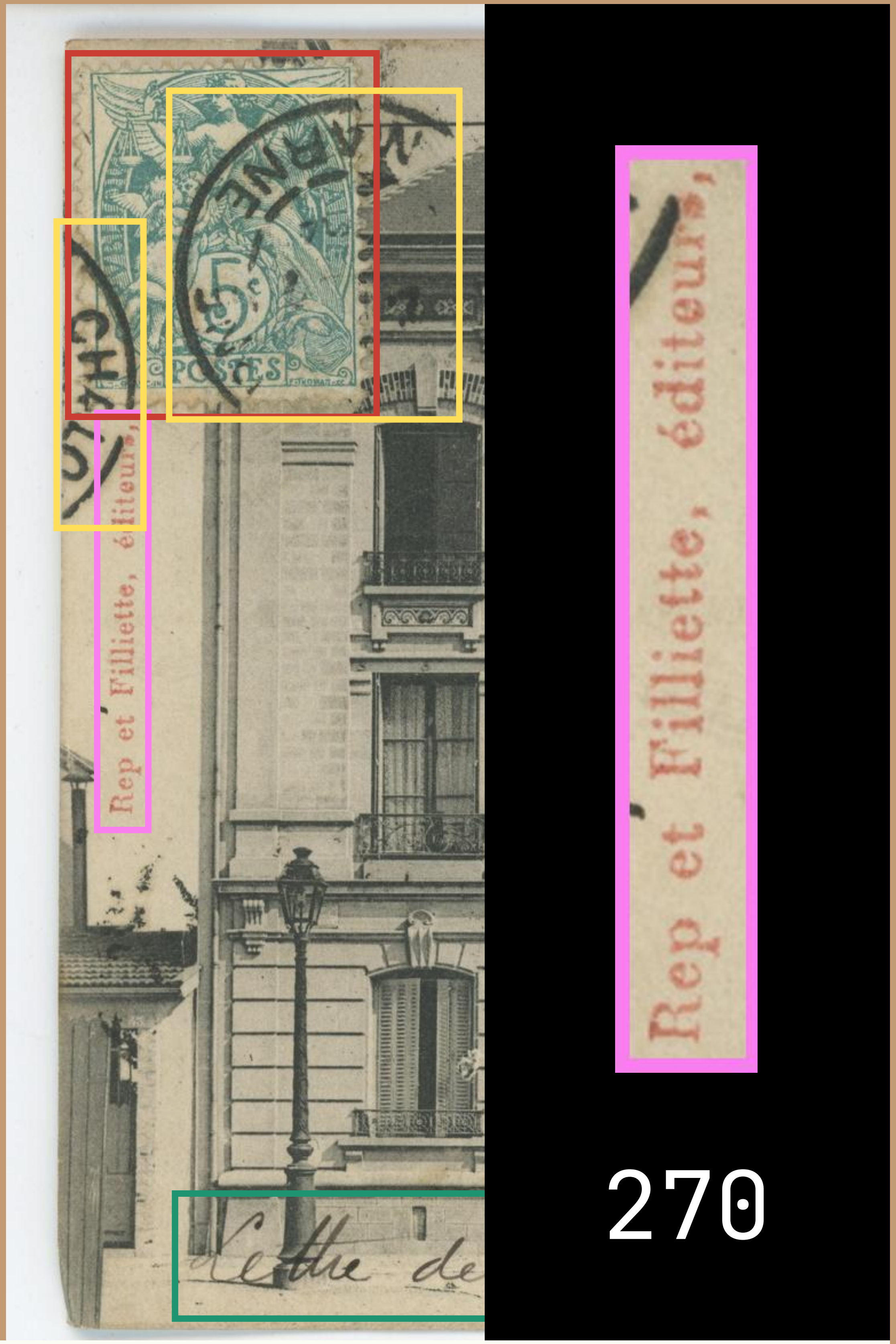
6 classes from **2 main categories**:

- **texts:** printed, handwritten or scene text
- **postal markings:** postage stamps, date stamps, other marks

=> Prevent interaction & allow ≠ processing methods

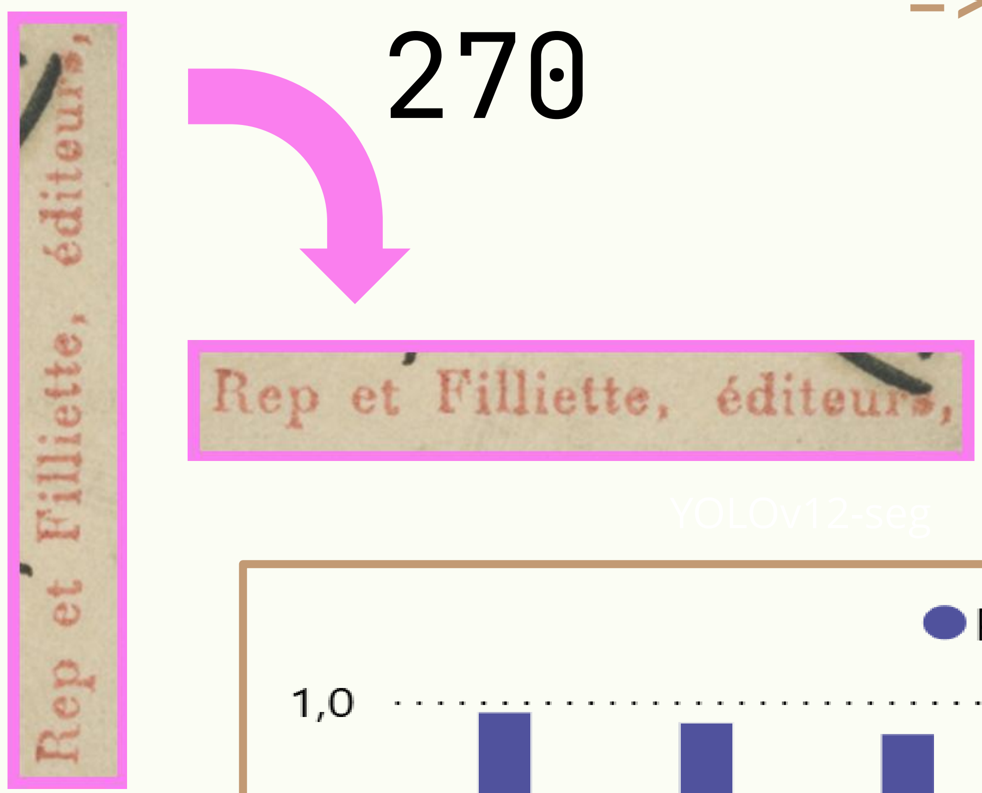
Scores (P, R & mAP@50):



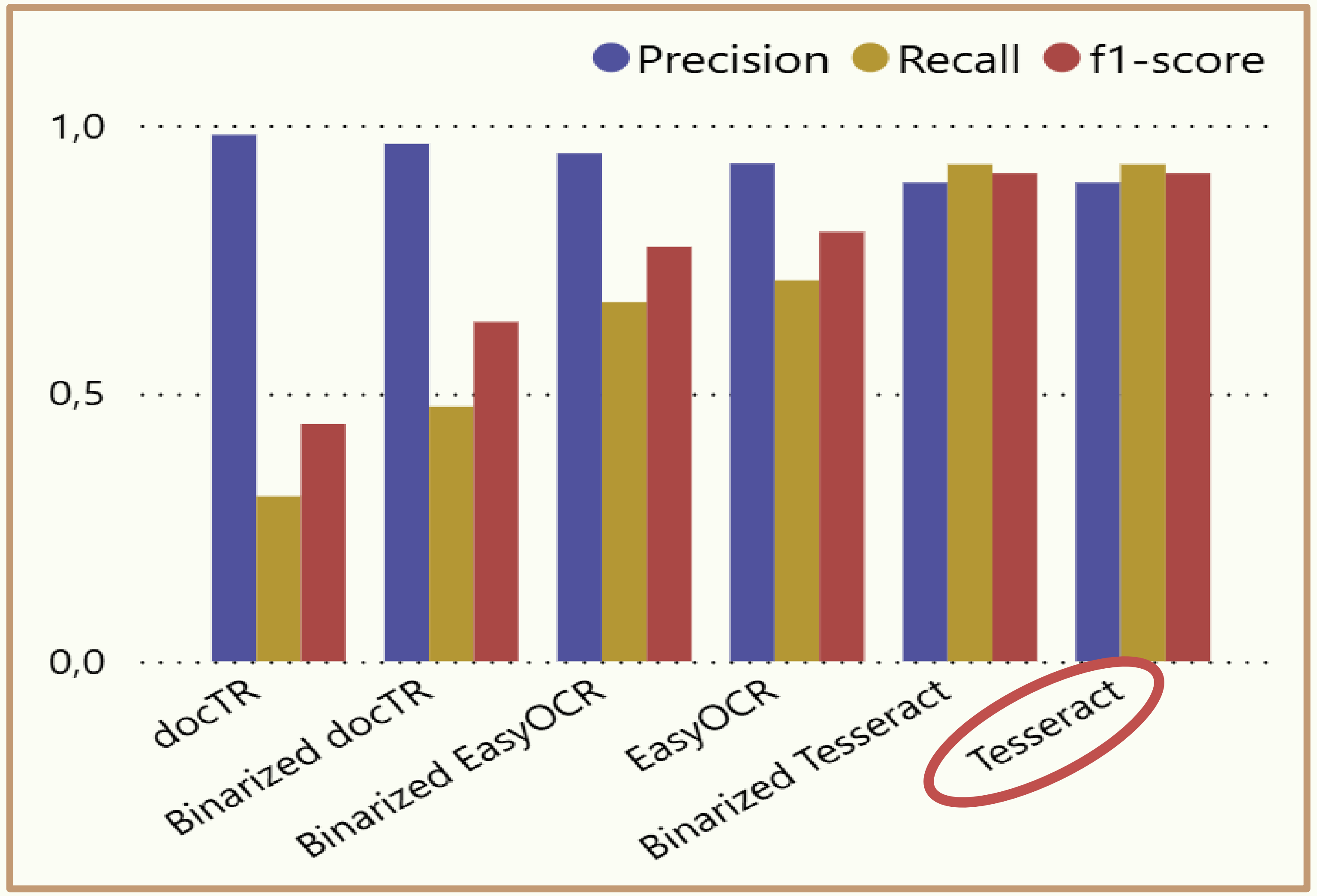


STEP 2: ORIENTATION

ORIENTATION PREDICTION

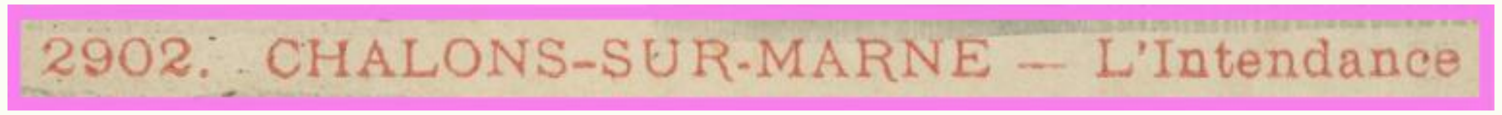


=> Most OCR algo read
0°-oriented text only



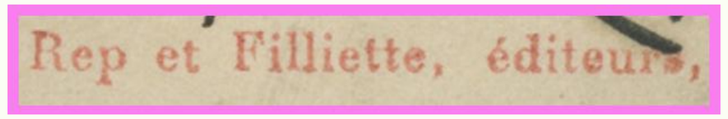
STEP 3: PRINTED TEXT TRANSCRIPTION

Transcription



2902. CHALONS-SUR-MARNE — L'Intendance

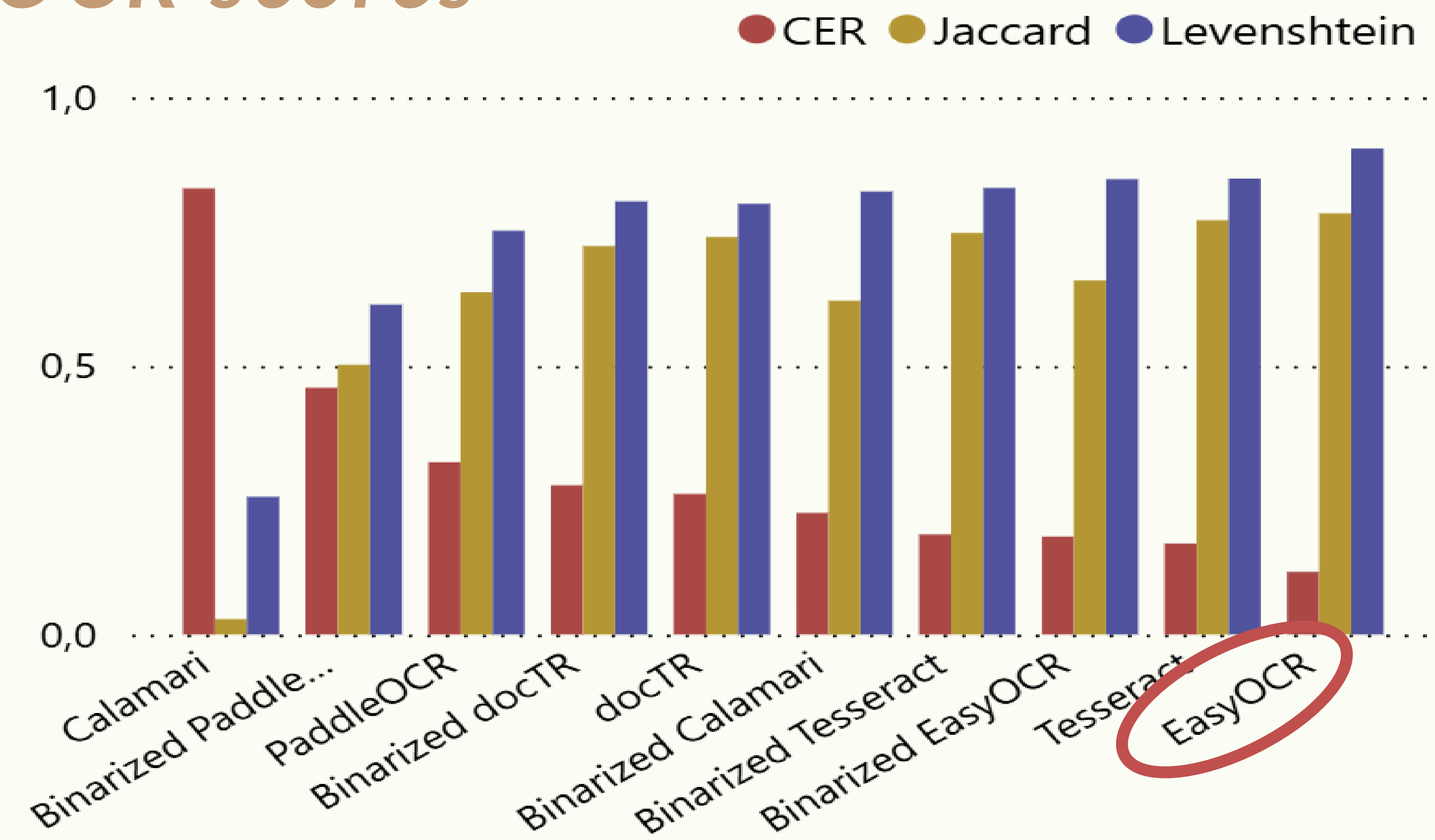
Printed 1: 2902. CHALONS-SUR-MARNE – L'Intendance



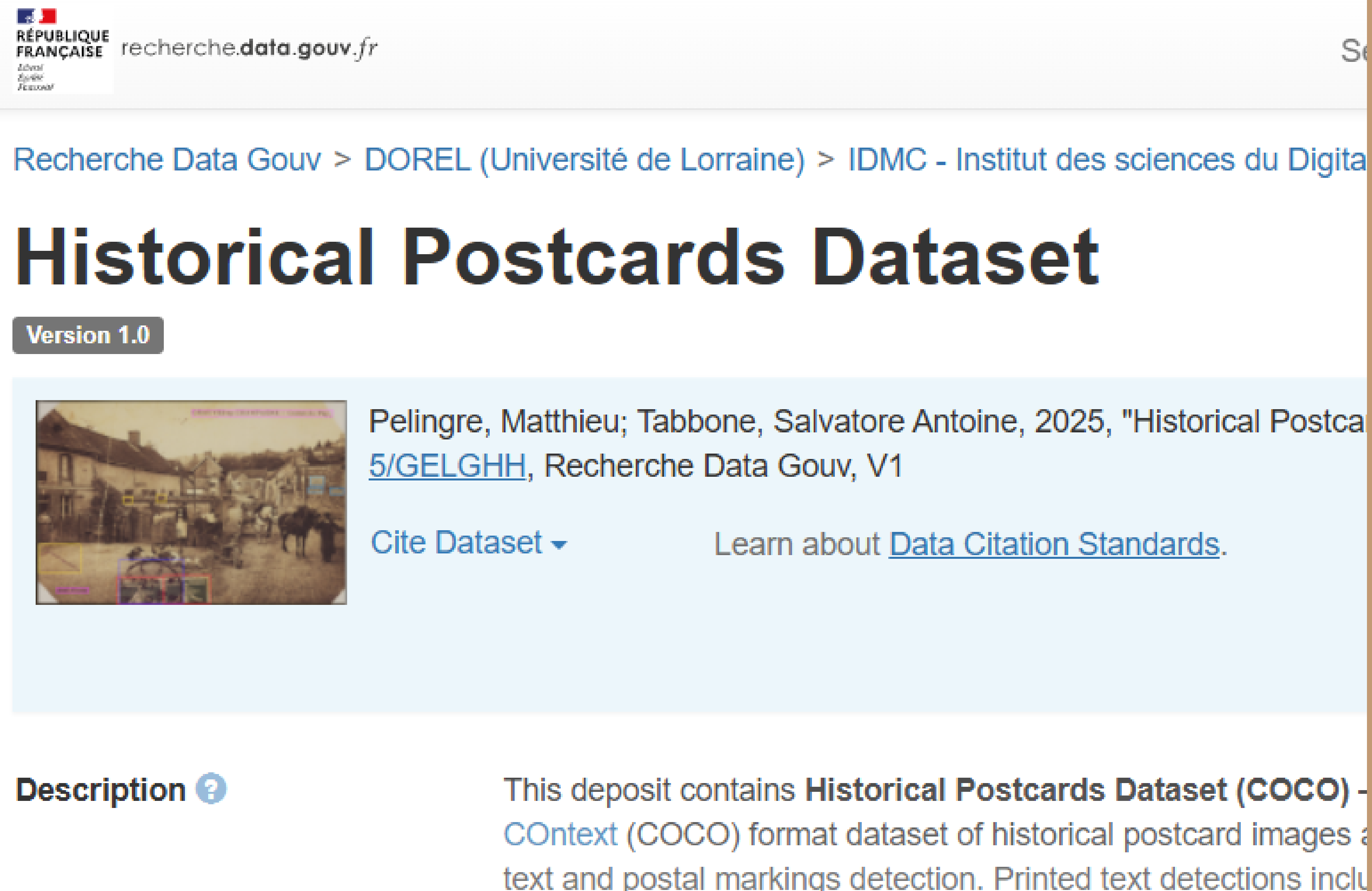
Rep et Filliette, éditeurs,

Printed 2: Rep et Filliette, éditeurs,

OCR scores



RECHERCHE.DATA.GOUV.FR



RÉPUBLIQUE FRANÇAISE recherche.data.gouv.fr

Se

Recherche Data Gouv > DOREL (Université de Lorraine) > IDMC - Institut des sciences du Digita

Historical Postcards Dataset

Version 1.0

Pelingre, Matthieu; Tabbone, Salvatore Antoine, 2025, "Historical Postca
5/GELGHH, Recherche Data Gouv, V1

Cite Dataset ▾

Learn about Data Citation Standards.

Description

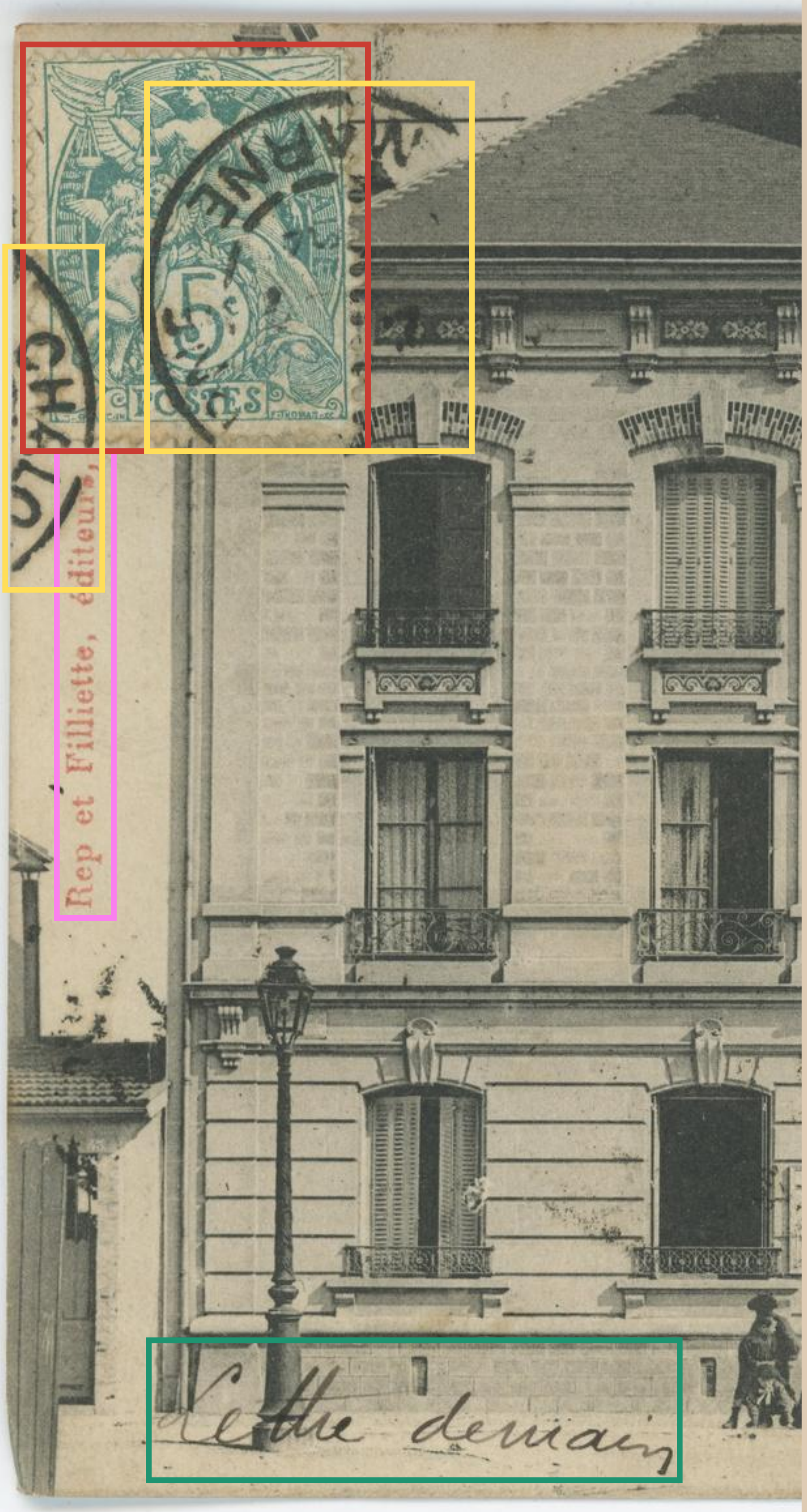
This deposit contains **Historical Postcards Dataset (COCO) -**
COntext (COCO) format dataset of historical postcard images a
text and postal markings detection. Printed text detections inclu

Log In

THE DATASET

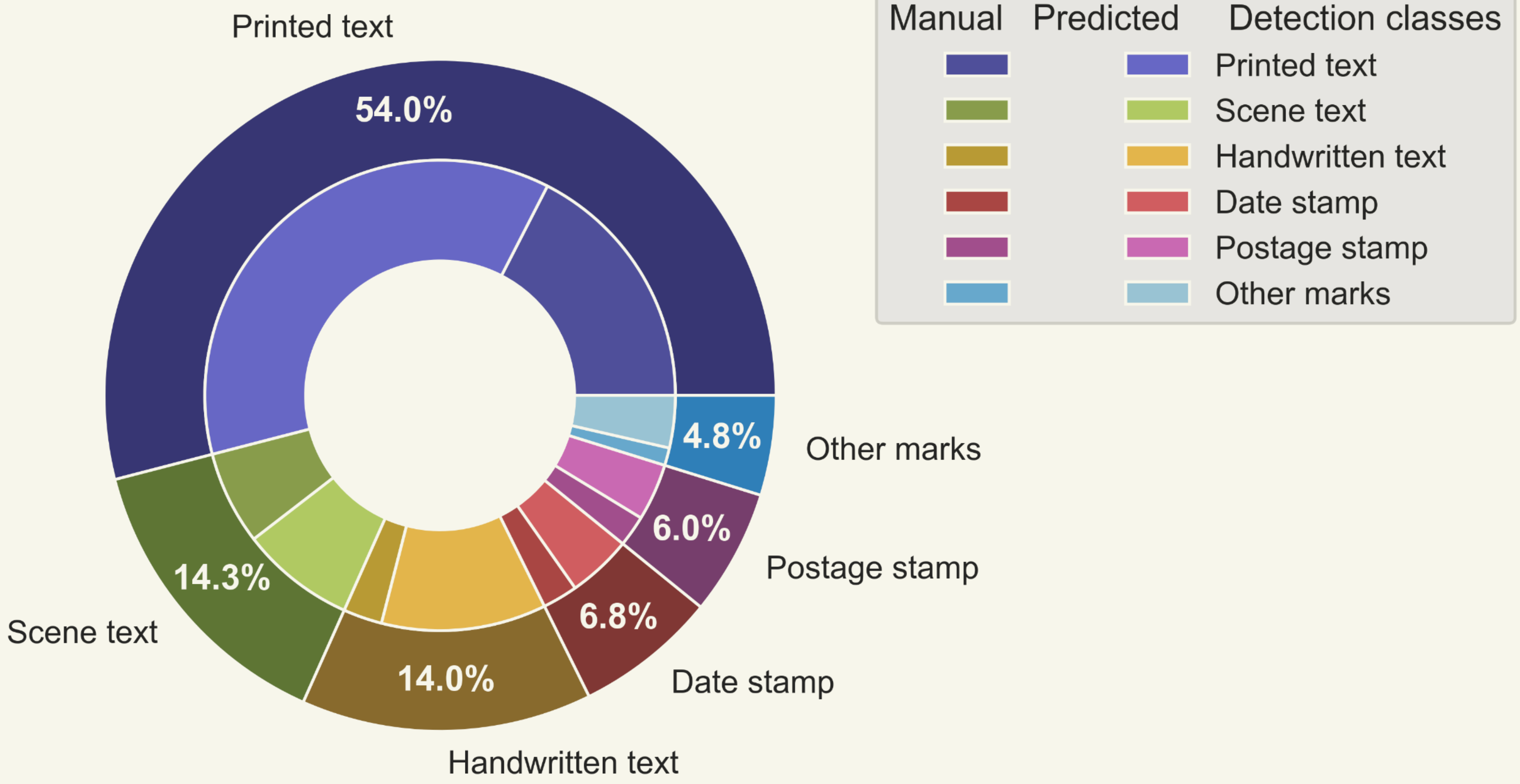
AVAILABLE ONLINE

- COCO-format
  - ✓ 6 detection classes
  - ✓ With orientation and transcription (for printed text)
  - ✓ Compatible with COCO-API & CVAT
- 3 subsets
  - ✓ Train and Test sets used in article
  - ✓ 1,375 manually annotated postcards
  - ✓ Synthetic set with 2,918 postcards



DATASET DESCRIPTION

DETECTIONS



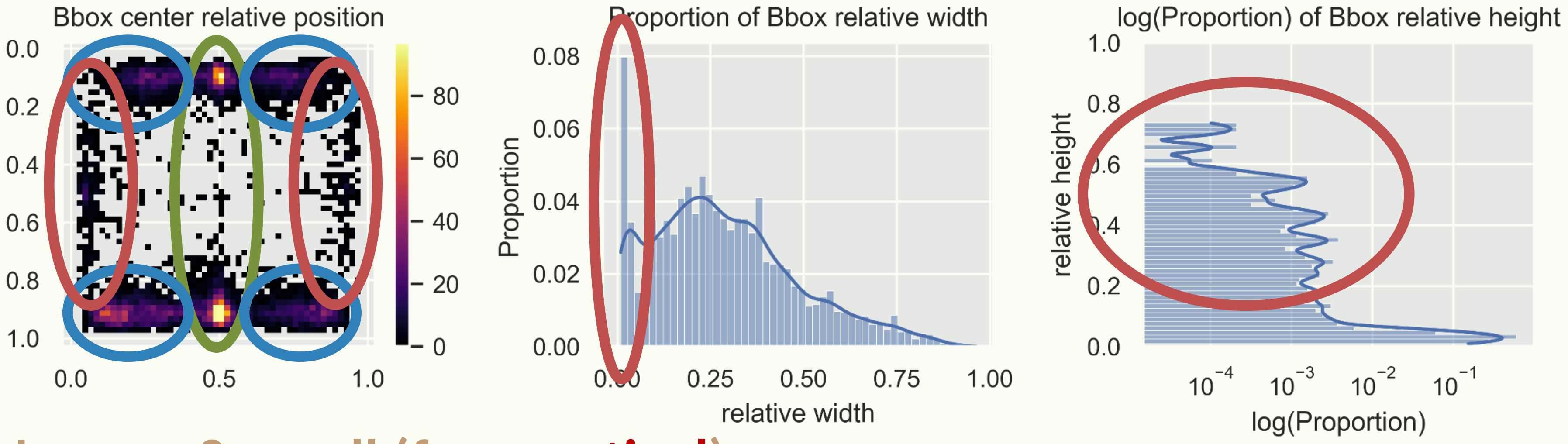
17,423 Detection Zones

Wrong ratio for **handwritten**, **scene text** & **other marks** :

variability, under-representation or **insufficient model input size**.

DATASET DESCRIPTION

PRINTED TEXT ZONES

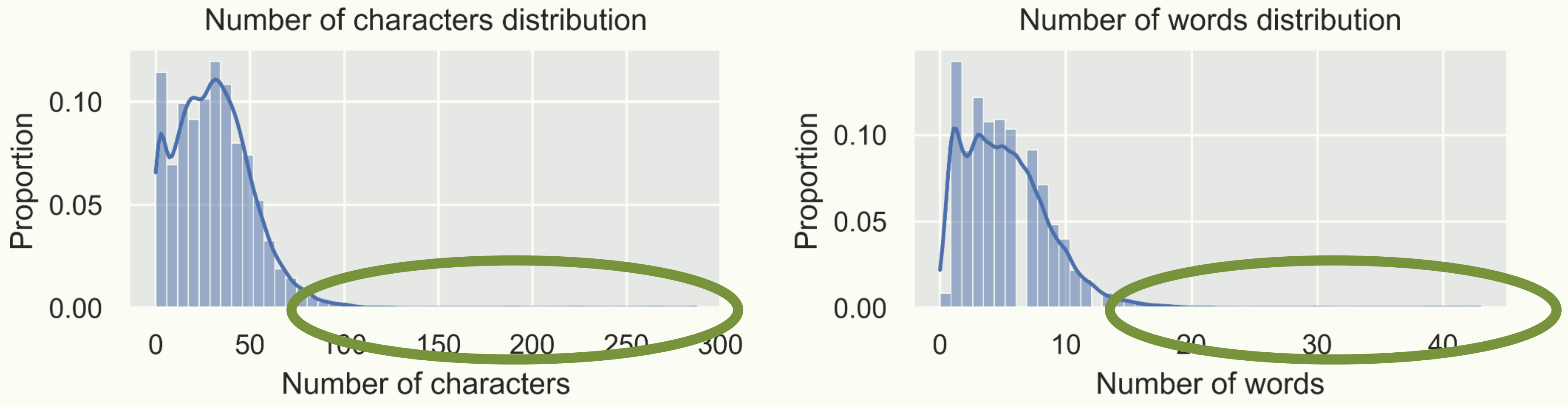


Narrow & small (few **vertical**)

Captions: middle top or bottom

Publisher information: top or bottom margins

TRANSCRIPTION



Concise & highly variable (short captions, publisher imprints)

Occasional : longer narrative or poems.

WHAT'S NEXT ?

FUTURE UPDATES

Printed Text:

- *Name Entity Recognition*
- *Keywords*

Date Stamps:

- *Segmentation/binarization*
- *Transcription*

Longer term:

- *More Postcards*
- *Handwritten Text transcription*
- *Scene Text transcription*

THANK YOU!

DATASET REPOSITORY:

DOI: 10.57745/GELGHH

ACKNOWLEDGMENTS

IMAGE'EST

La Région Grand Est