



**HAL**  
open science

# CREMMALab Project: Handwritten Text Recognition for medieval manuscripts

Ariane Pinche

► **To cite this version:**

Ariane Pinche. CREMMALab Project: Handwritten Text Recognition for medieval manuscripts. Digital Humanities, Jul 2022, Tokyo, Japan. hal-03724041

**HAL Id: hal-03724041**

**<https://hal.science/hal-03724041v1>**

Submitted on 15 Jul 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# CREMMALAB Project

## Handwritten Text Recognition (HTR) for medieval manuscripts

Within the infrastructure of the CREMMA project (Consortium for Handwriting Recognition of Ancient Materials) supported by the DIM (research funded by the Île-de-France Region) MAP (Ancient and Heritage Materials), the CREMMALab project (2021-2022) combines research questions, creation and release of data from French medieval literary manuscripts for HTR.

Project Manager:

Ariane PINCHE (postdoctoral researcher)

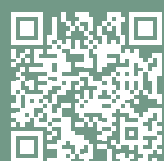
### EVENTS

Conference *Ancient documents and automatic recognition of handwriting*, 23-24 June 2022

### WEBSITES



<https://cremmalab.hypotheses.org>



<https://github.com/HTR-United/cremma-medieval>

## CREMMA MEDIEVAL DATASET

Medieval manuscripts corpus for training HTR models from the 12<sup>th</sup> to the 15<sup>th</sup> century, included in HTR-United catalog: <https://htr-united.github.io>

Dataset has been built with eScriptorium, an interface for HTR ground truth production, and Kraken, an HTR and layout segmentation engine. It is composed of fourteen Old French manuscripts written between the 13<sup>th</sup> and 15<sup>th</sup> centuries.

Transcriptions have been standardized to strengthen HTR models. We chose a graphemic transcription method. Transcription guidelines available here: <https://hal.archives-ouvertes.fr/hal-03697382>

MANUSCRIPT	DATE	TRANSCRIBED LINES
BnF, ms fr. 412	13th	6324
BnF, Arsenal 3516	13th	1991
Cologne, bodmer, 168	13th	1976
BnF, ms fr. 24428	13th	1328
BnF, ms fr. 25516	13th	717
BnF, ms fr. 844	13th	224
BnF, ms fr. 17229	13th	164
BnF, ms fr. 13496	13th	161
BnF, Arsenal 3516	13th	105
BnF, ms fr. 22549	14th	2682
Vaticane, Reg. Lat., 1616	14th	1772
Univ. of Pennsylvania, 660	14th	368
BnF, ms fr. 411	14th	179
Univ. of Pennsylvania, 909	15th	2513
All		21656

To ensure data quality, continuous integration workflow has been put in place:

- HTRVX: XML schema validator checking Alto files and segmentation vocabulary (segmOnto);
- Choco-Mufin: checking the homogeneity of the characters used in the dataset.

Abstract: <https://hal.archives-ouvertes.fr/hal-03719504>

## HTR MODELS FOR MEDIEVAL MANUSCRIPTS TRAINED WITH KRAKEN

<https://kraken.re>

### Test scores in-domain

**1.0.1 Bicerin**, accuracy 89.19%, model based on CREMMA Medieval first dataset, specialized on 13<sup>th</sup> and 14<sup>th</sup> c. manuscripts.

**1.1.0 Bicerin**, accuracy 95.30%, model based on CREMMA Medieval extended to 15<sup>th</sup> c. manuscripts.

**2.0.0 Cortado**, accuracy 95.54%, model mixing CREMMA Medieval dataset with early prints (15<sup>th</sup> c.) from Gallic(orpor)a Project.

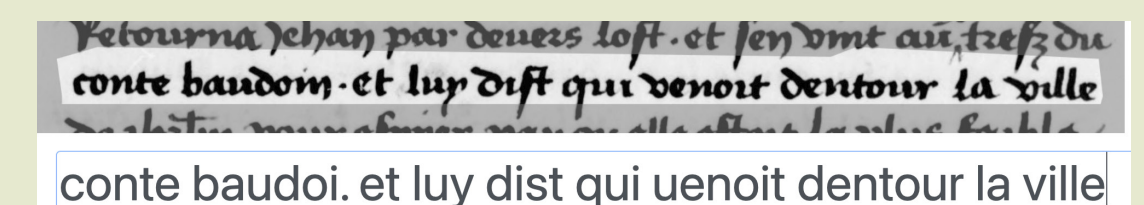
Models available here:

<https://github.com/HTR-United/cremma-medieval/releases>

### Test scores out-of-domain

	BnF, ms, fr. 17229, 13th c.	BnF, ms, fr. 185, 14th c.	BnF, NAF 6213, 15th c.	ALL
Cortado	92.71%	92.07%	87.48%	90.95%
1.1.0 Bicerin	91.64%	91.34%	83.40%	89.23%
1.0.1 Bicerin	90.66%	88.45%	79.67%	86.50%

### Example, Cortado model on BnF, NAF 6213



Conclusion: a specialized model per script isn't always necessary, but the variety of the training set increases its robustness, even in our case with early prints.

See complete experience here:

<https://cremmalab.hypotheses.org/modeles-htr>

