[1]CIHAM (UMR 5648), CNRS, [2]ALMAnaCH, INRIA

CATMuS-Medieval

Consistent Approaches to Transcribing ManuScripts

Authors : A. Pinche, T. Clérice, A. Chagué, J-B. Camps, M. Vlachou-Efstathiou, M. Gille Levenson, O. Brisville-Fertin, F. Boschetti, F. Fischer, M. Gervers, A. Boutreux, A. Manton, S. Gabay, P. O'Connor, W. Haverals, M. Kestemont, C. Vandyck

Presenting speakers: Ariane Pinche[1], Thibault Clérice[2]
ariane.pinche@cnrs.fr, thibault.clérice@inria.fr

DH 2024: Reinvention & Responsibility, August 8, 2024

Table of Contents

Automatic Text Recognition: State of the Art

- **2016**, the OCRopy tool (Breuel 2014)– designed as an open-source, easily trainable OCR engine – piqued the interest of digital humanists
- **2016–2019**, the landscape of handwritten text recognition (HTR) underwent a profound transformation
  - Model training became feasible with the advent of general public GPUs
  - Programming libraries like Tensorflow and PyTorch became standard
  - Kraken gained prominence (Kiessling 2022)
  - Transkribus emerged as a comprehensive solution, enabling manual annotation, model training, and post-correction, (Kahle, Colutto, Hackl, and Mühlberger 2017)

Automatic Text Recognition (ATR): State of the

- **2020 - ...**, ATR became more and more common in research projects in humanities
  - Expansion of the use of intuitive platforms: eScriptorium, (Stokes et al. 2021) and Transkribus
  - Need to standardize methods and data sets to enable data gathering and production of a generic model

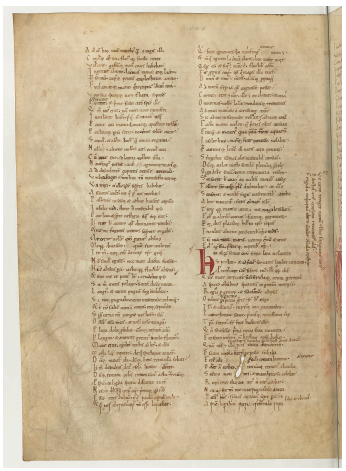


Figure: BnF, Latin, 8001, 13th century

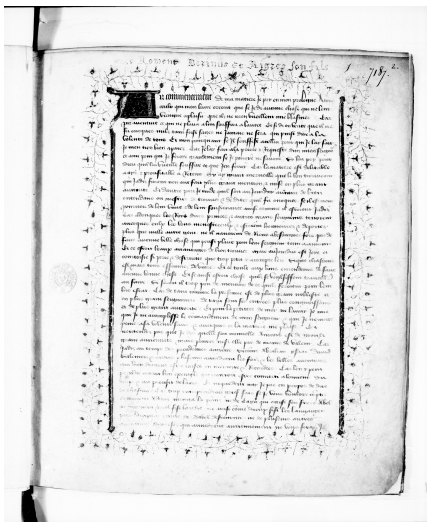


Figure: BnF, French, 777, 15th century

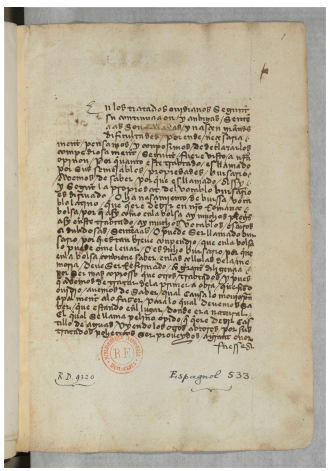


Figure: BnF, Espagnol, 533, 15th century

CATMuS Project Overview

- **CATMuS** stands for **Consistent Approaches to Transcribing ManuScripts**.
- International initiative involving 17 collaborators: A. Pinche, T. Clérice, A. Chagué, J-B. Camps, M. Vlachou-Efstathiou, M. Gille Levenson, O. Brisville-Fertin, F. Boschetti, F. Fischer, M. Gervers, A. Boutreux, A. Manton, S. Gabay, P. O'Connor, W. Haverals, M. Kestemont, C. Vandyck
- From Europe and North America
- Provides guidelines, datasets, and models for HTR

Table of Contents

Once upon a time DH 2019...

Everything began at DH 2019, with: Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. “Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer’s hagiographic hypothesis”. In: *Digital Scholarship in the Humanities* 36.Supplement_2 (Oct. 2021), pp. ii49–ii71. ISSN: 2055-7671. URL: https://doi.org/10.1093/llc/fqab033 (visited on 12/03/2021)

CREMMA & CREMMALab

Collective initiatives on HTR, in partnership with École nationale des chartes (ENC), DIM MAP and INRIA:

- CREMMA (Consortium Reconnaissance d'Écriture Manuscrite des Matériaux Anciens) to create HTR datasets and help set up an eScriptorium instance
- CREMMALab
  - Methodological reflection on corpus transcription protocols to optimizing HTR models for medieval French Manuscripts
  - Producing a dataset to train generic model for medieval French manuscripts, (Pinche 2022)
- CREMMA Medii Aevi: dataset for medieval manuscripts in Latin, (Clérice, Chagué, and Vlachou-Efstathiou 2023)

CATMuS Chronology

1. CREMMA & CREMMALab (DIM MAP, région île de France)
2. Gallic(orpor)a
3. FoNDUE
4. DEEDS
5. HTRomance (training data production for French and Latin manuscripts, Bibliothèque nationale de France)
6. HTRogène (multilingual training data production, Biblissima+, ongoing)

Total expenses: €100,000 for infrastructure, postdoctorat, training data producing, not counting “hours” of team members.

Table of Contents

Why establishing transcription guidelines ?

> *"Well-prepared material is key to producing general recognition models. It is unthinkable that single scholars and small project teams could provide enough training material to train a general model independently." (Hodel, Schoch, Schneider, and Purcell 2021)*

Transcription guidelines are needed to:

- Create **consistent data**
- Build **reusable ground truth data** sets to minimize the collective cost
- Produce **generic models**

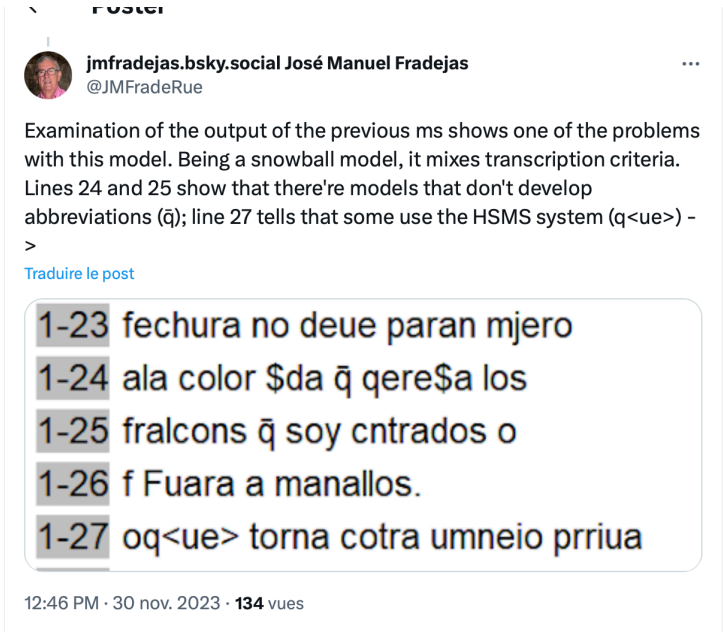


Figure: Prediction using Transkribus Coloso Español model. Capture kindly provided by J. M. Fradejas Rueda.

Birth of the CATMuS-Medieval Guidelines

- The CATMuS guidelines follow on from those of the CREMMALab project, *available here: https://hal.archives-ouvertes.fr/hal-03697382*
- This first draft is the results of a reflection carried during the years 2021 and 2022 at the ENC.
- The aim was:
  - To ensure compatibility of transcription data
  - To find a medium path to ensure feasibility and consistency of the data
  - To produce data reusable by historians, linguists, paleographists philologists
  - Not to constitute a definitive transcription or a final edition

Figure: Example of a conceptual text acquisition pipeline

CATMuS-Medieval guidelines: principles

How to transcribe manuscripts to ensure consistency in the HTR model?

- Define transcription methods suitable for machine learning.
- Determine the desired level of detail in transcription.

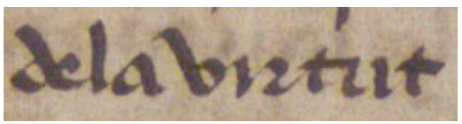


Figure: ‹de la uirtut› or ‹ɗela ʋɪɾtut›

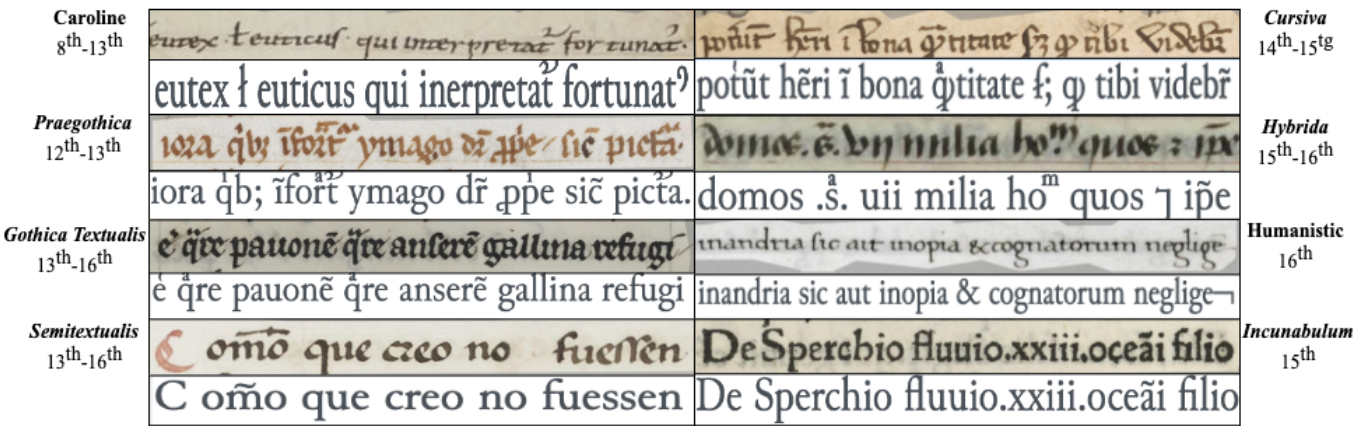
- Ensure compatibility of data
  - Use a predefined characters set
  - Document your choices

LATIN SMALL LETTER P WITH STROKE (*p barré droit*)		ꝑ	U+A751
LATIN SMALL LETTER P WITH FLOURISH (*p barré courbe*)		ꝓ	U+A753
LATIN SMALL LETTER Q WITH DIAGONAL STROKE (*q barré*)		ꝗ	U+A759
TIRONIAN SIGN ET (*abréviation tironienne de « et »*)		⁊	U+204A
DIVISION SIGN (*abréviation de « est »*)		÷	U+00F7

CATMuS-Medieval guidelines: principles

Guidelines are based on basic principles:

- Reducing each letter form to a standardized representation
- Distinctions like ‹u›/‹v› and ‹i›/‹j› are reduced to ‹u› and ‹i›
- Keeping abbreviations (and thus, reducing the part of the language-specific traits the model has to learn)
- Restricting the characters used for specific purposes according to the MUFI
- Reducing the complexity of medieval punctuation: single sign = “.” and double sign = “:”.



Ongoing efforts

- New languages and new challenges:
  - Latin involves new abbreviation to represent, such as <ꝛ> for <rum>;
  - Middle English involves new signs for new phonetic realization, like <ð> (eth) representing the voiced and voiceless dental fricative.
- The establishment of transcription rules remains an ongoing and evolving effort for ensuring data quality and homogeneity.
- The guidelines will be available at the following link: `https://catmus-guidelines.github.io` - Work in progress.

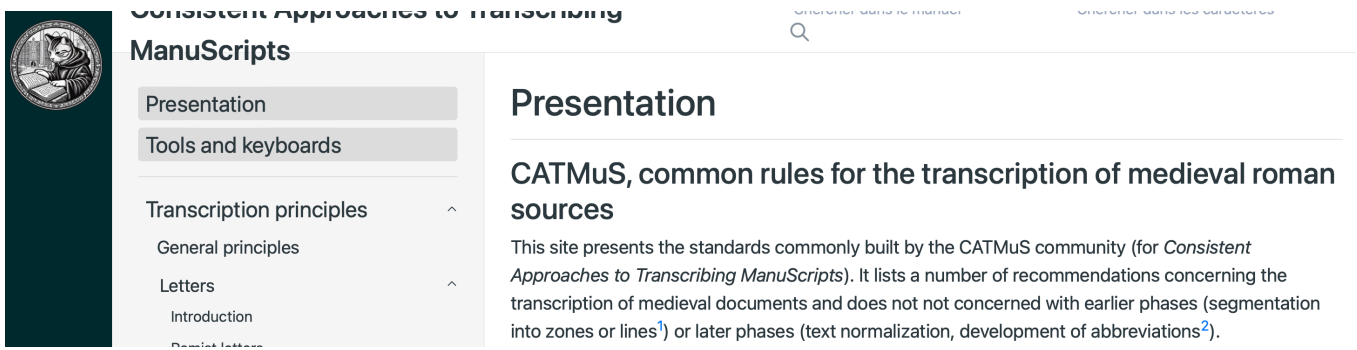
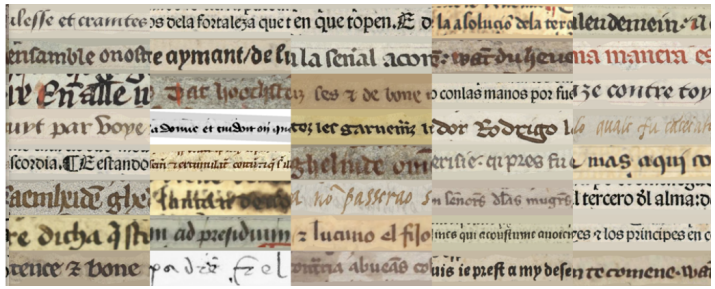


Table of Contents

CATMuS-Medieval dataset

CATMuS Medieval



- CATMuS dataset is published and documented on hugging face: `https://huggingface.co/datasets/CATMuS/medieval`
- built upon 17 different repositories
- c. 175.000 lines and 6.2M characters
- contains 245 different documents
- 9 different langages
- Its most represented centuries are the 15th century, the 14th, and the 13th.
- Over-representation of some genre and language: French-Narratives (29.2% of the dataset), Treatises-Latin (24.2%) and Treatises-Castillan (23%), linked to the history of the data gathering

CATMuS-Medieval models

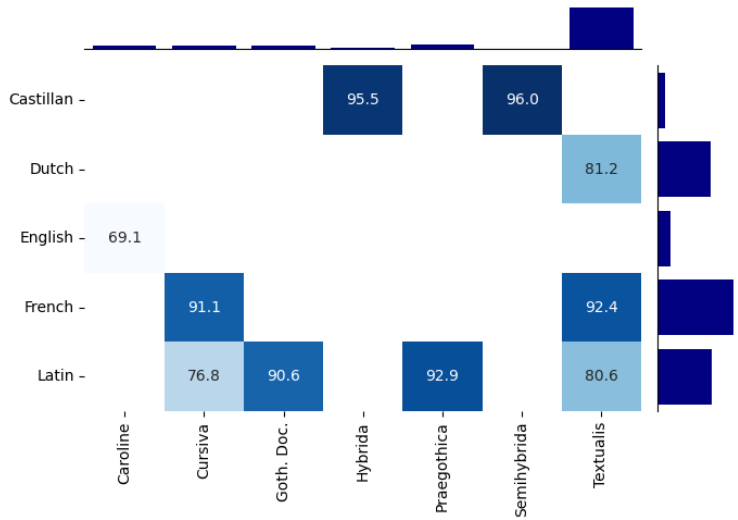


Figure: Absolute accuracy

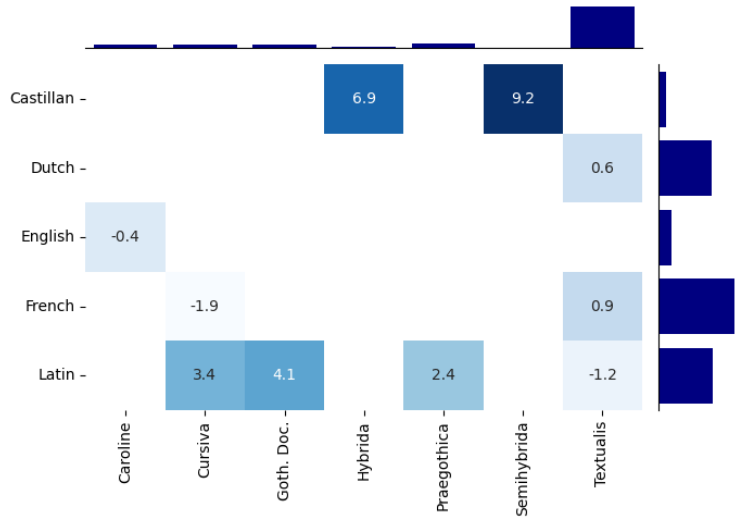


Figure: Improvement over CREMMA Generic

Figure: Test micro-accuracy per language and scripta. Bars represents the total number of characters per categorical feature.

CATMuS Generic Models

- A general model for medieval manuscripts:Ariane Pinche et al. “CATMuS Medieval”. lat. In: (Nov. 2023). Publisher: Zenodo. URL: https://zenodo.org/records/10066219 (visited on 01/08/2024)
- A general model for gothic prints:Sonia Solfrini and Simon Gabay. “CATMuS Gothic Print”. frm. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10599911 (visited on 03/27/2024)
- A general model for prints: Simon Gabay and Thibault Clérice. “CATMuS-Print [Large]”. fra. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10592716 (visited on 03/27/2024)

Table of Contents

Conclusion

- Thanks to the guidance provided by the guidelines and the support of our sponsors, we have significantly **increased the number of available, interoperable datasets** for training HTR models in recent years.
- This increase has enabled us to develop **new generic models for medieval manuscripts**.
- Consistent transcription standards, collaboration and data sharing across languages have **improved the accuracy of HTR models**
- *Future developments:*
  - Continuing the extension of the project to Old English and Middle Dutch
  - Extending the temporal scope from the Middle Ages to the contemporary era

References I

[1] Thomas M. Breuel. *Ocropy: Python-based tools for document analysis and OCR*. 2014. URL: https://github.com/tmbdev/ocropy..

[2] Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. "Noisy medieval data, from digitized manuscript to stylometric analysis: Evaluating Paul Meyer's hagiographic hypothesis". In: *Digital Scholarship in the Humanities* 36.Supplement_2 (Oct. 2021), pp. ii49–ii71. ISSN: 2055-7671. URL: https://doi.org/10.1093/llc/fqab033 (visited on 12/03/2021).

[3] Thibault Clérice, Alix Chagué, and Malamatenia Vlachou-Efstathiou. *CREMMA Medii Aevi*. 2023. URL: https://github.com/HTR-United/CREMMA-Medieval-LAT.

[4] Simon Gabay and Thibault Clérice. "CATMuS-Print [Large]". fra. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10592716 (visited on 03/27/2024).

[5] Tobias Mathias Hodel, David Selim Schoch, Christa Schneider, and Jake Purcell. "General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example". In: *Journal of open humanities data* 7.13 (2021), pp. 1–10.

[6] Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 04. ISSN: 2379-2140. Nov. 2017, pp. 19–24. DOI: 10.1109/ICDAR.2017.307.

References II

[7] Benjamin Kiessling. *The Kraken OCR system*. Version 4.1.2. Apr. 2022. URL: https://kraken.re.

[8] Ariane Pinche. *Cremma Medieval*. June 2022. URL: https://github.com/HTR-United/cremma-medieval.

[9] Ariane Pinche et al. "CATMuS Medieval". lat. In: (Nov. 2023). Publisher: Zenodo. URL: https://zenodo.org/records/10066219 (visited on 01/08/2024).

[10] Sonia Solfrini and Simon Gabay. "CATMuS Gothic Print". frm. In: (Jan. 2024). Publisher: Zenodo. URL: https://zenodo.org/records/10599911 (visited on 03/27/2024).

[11] Peter A. Stokes, Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, and El Hassane Gargem. *The eScriptorium VRE for Manuscript Cultures – Classics@ Journal*. en-US. 2021. URL: https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/ (visited on 12/06/2023).