



# HTR-United Workshop

Metadata, quality control and sharing process for  
HTR training data

**Alix Chagué** (ALMAAnACH, UdeM, ÉPHE)

**Thibault Clérice** (ALMAAnACH, CJM)

**Hugo Scheithauer** (ALMAAnACH)

July 12 2023

DH2023, Graz, Austria





## Plan of the workshop

Or what we are gonna do today

### (A) Slides

- Recap' on HTR and Layout segmentation
- What's HTR-United
- Virtual guided tour

### (C) About the *hands-on* (the right column)

Everytime we feel the need, we stop, we discuss. If we don't discuss thing during the session, we'll talk about it after

### (B) Then... hands-on + discussion

- Accumulating data
- Organizing data
- Writing a README.MD
- Write a CITATION.CFF
- Put the dataset online (github, zenodo, both)
- Create the HTR-United.yml
- Offer the dataset on HTR-United
- Integration tools
- Publish HTR-United (We demonstrate)



# Table of Contents

1 Introduction

▶ Introduction

▶ What is HTR-United?

▶ Why contribute?



# Quick recap: from image to text, the steps

## 1 Introduction

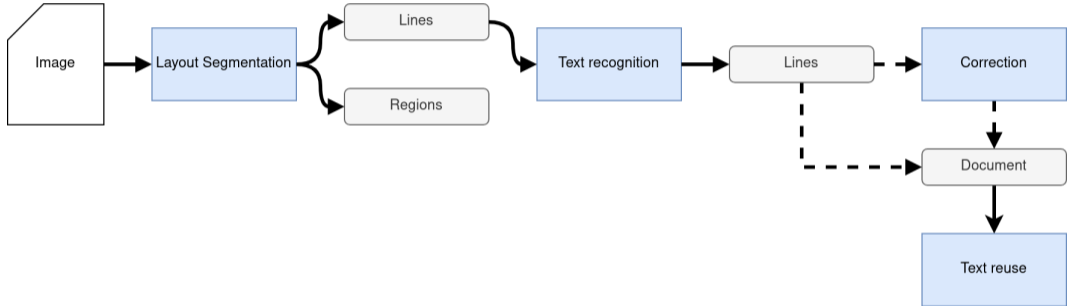


Figure: Different steps associated with retrieving the text of an image

# Layout Segmentation

## 1 Introduction

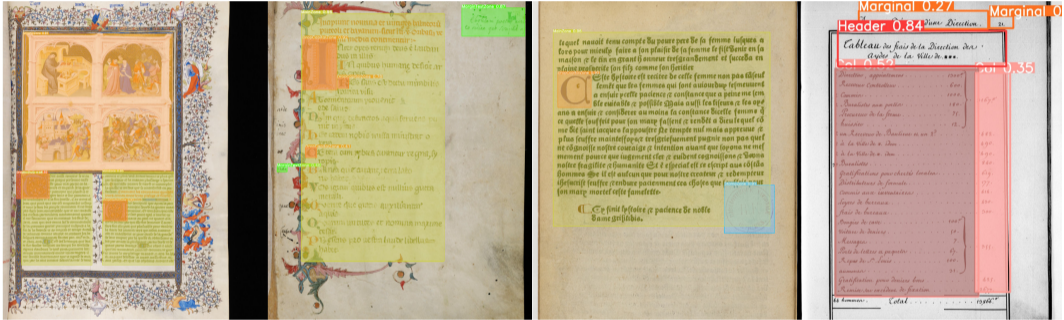
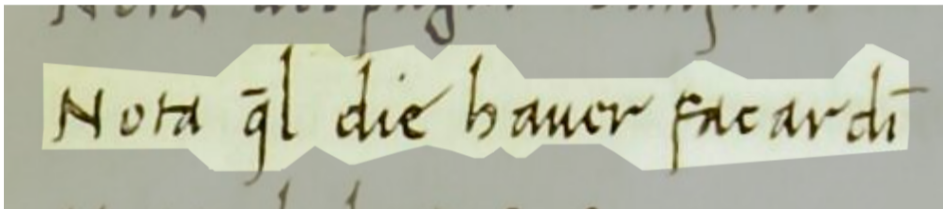


Figure: Examples of layout segmentation at the zone level




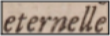



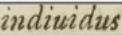


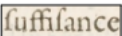



Nota q̄l die haver faeardī

Figure: Example of a corrected line

# Transcription guidelines ?

## 1 Introduction

Category	Description	Status	Transcription	Example
Ligature	Ligature O+E <œ>	Graphetic	U+0153/U+0152	
Ligature	Ligature A+E <æ>	Graphetic	U+00E6/U+00C6	
Ligature	Ligature long S+T <ft>	Graphemic	No ligature	
Ligature	Ligature L+L <ll>	Graphemic	No ligature	
Ligature	Ligature C+T <ct>	Graphemic	No ligature	
Ligature	Ligature S+P <sp>	Graphemic	No ligature	
Ligature	Ligature long S+L <fl>	Graphemic	No ligature	
Ligature	Ligature U+S <us>	Graphemic	No ligature	
Ligature	Ligature S+I <fi>	Graphemic	No ligature	
Ligature	Ligature long S+long S <ff>	Graphemic	No ligature	
Ligature	Ligature F+F+I <ffi>	Graphemic	No ligature	
				

# Training and fine-tuning

## 1 Introduction

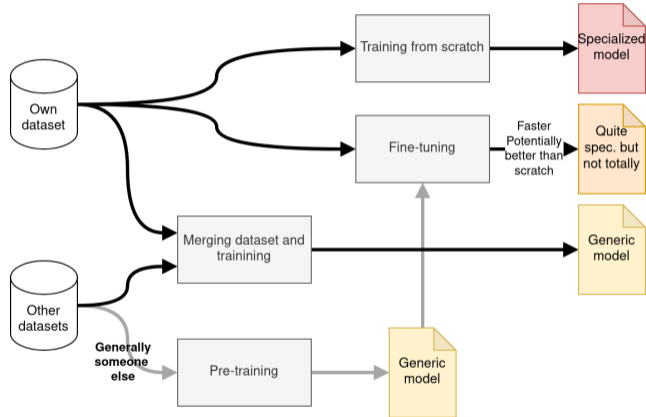


Figure: Example of workflows





# Situation with HTR and HTR Data

## 1 Introduction

- OCR and HTR are great opportunities to access collections of documents and create textual corpora
- but transcription models are costly to produce because they require training examples
- we need to rely on pre-existing models and pre-existing data
- they are rarely FAIR
  - hard to **F**ind and not always **A**ccessible
  - uncertain formats & varying annotations
  - unclear **R**euse conditions



# Table of Contents

## 2 What is HTR-United?

▶ Introduction

▶ What is HTR-United?

▶ Why contribute?



## It's a catalog

2 What is HTR-United?

Some of our guiding principles:

- Browsable for humans thanks to a user interface offering filters
- Actionable for machines through a structured, documented and versioned catalog synched with Zenodo
- A low tech environment to insure it is easy for us to maintain it



## It's a catalog

### 2 What is HTR-United?

The catalog is fed by the creators of the datasets. How do they contribute?

- data publication (we offer a template and guidelines for good practices)
- creation of the new catalog entry ('htr-united.yml') using our form
- interaction through Github issues to fix issues we spot while validating the entry



## It's a catalog

### 2 What is HTR-United?

As of July 4th, 2023, the catalog refers to:

- 78 datasets created by at least 36 different projects
- 21 languages (a lot of French, and Latin) for 7 scripts (mostly Latin)
- handwritten and printed documents, mixed or "pure"
- a period going from 800 to 2023
- created with at least 6 different HTR software
- more than 44M characters, than 1M lines, or than 20K images



## It's a schema

### 2 What is HTR-United?

- controlled vocabulary formalized using JSONSchema
- contains metadata such as:
  - desc. of the ground truth (language and script, number of hands, period covered, character set)
  - desc. of the dataset (link, title, desc., file format, metrics, licence)
  - desc. of the condition of production (project, authors and annotators, software)
- longer term goal: building a controlled vocabulary for transcription guidelines
- its evolution is transparent and documented through GitHub issues





## It's a toolbox

### 2 What is HTR-United?

Several actions are common to different projects so we created tools to help automatize them

- **HTRuc**: controls the validity of the htr-united.yml files and helps building the main catalog file
- **HTRVX**: controls the validity of the XML files (including to ontologies like SegmOnto and presence of empty elements)
- **HumGenerator**: computes metrics (files, regions, lines, chars), creates nice badges to display them, updates htr-united.yml
- **ChocoMufin**: controls chars in a dataset, converts chars according to a conversion table (*originally dev. for the CREMMA Medieval corpus by A. Pinche and T. Clérice*)





## Virtual tour of the place

2 What is HTR-United?

Let's go to `https://htr-united.github.io`



# Table of Contents

3 Why contribute?

▶ Introduction

▶ What is HTR-United?

▶ Why contribute?



# The stakes for the scientific community

## 3 Why contribute?

- HTR-United helps being FAIR
- it advocates for the recognition of dataset as scientific outcomes
- it helps creating (generic) models on a greater variety of data (because the data is shared)
- it paves a way for the standardisation of transcription practices across platforms, languages and scripts



## Plan of the workshop

Or what we are gonna do today

### (A) Slides

- Recap' on HTR and Layout segmentation
- What's HTR-United
- Virtual guided tour

### (C) About the *hands-on* (the right column)

Everytime we feel the need, we stop, we discuss. If we don't discuss thing during the session, we'll talk about it after

### (B) Then... hands-on + discussion

- Accumulating data
- Organizing data
- Writing a README.MD
- Write a CITATION.CFF
- Put the dataset online (github, zenodo, both)
- Create the HTR-United.yml
- Offer the dataset on HTR-United
- Integration tools
- Publish HTR-United (We demonstrate)