

HTR-United Workshop

Metadata, quality control and sharing process for HTR training data

Alix Chagué (ALMAnaCH, UdeM, ÉPHE) Thibault Clérice (ALMAnaCH, CJM) Hugo Scheithauer (ALMAnaCH)

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)



► Introduction

▶ What is HTR-United?

▶ Why contribute?



Quick recap: from image to text, the steps 1 Introduction

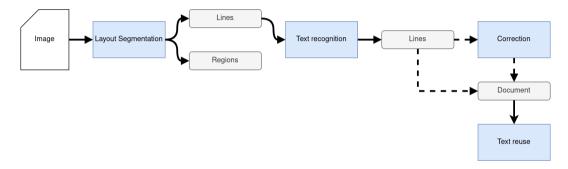


Figure: Different steps associated with retrieving the text of an image



Layout Segmentation

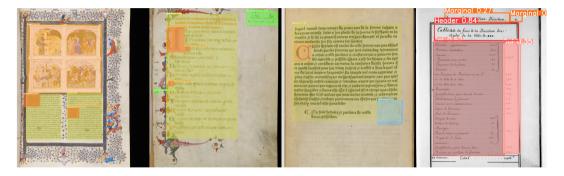
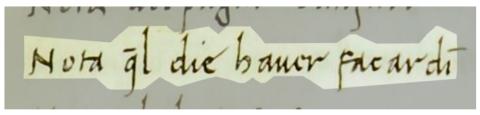


Figure: Examples of layout segmentation at the zone level





Nota ql die haver faeardi

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	cœur
Ligature	Ligature A+E <æ>	Graphetic	U+00E6/U+00C6	Cz- Ægyfte
Ligature	Ligature long S+T dt>	Graphemic	No ligature	Chrestienté.
Ligature	Ligature L+L dl>	Graphemic	No ligature	eternelle
Ligature	Ligature C+T «ct»	Graphemic	No ligature	Edict?
Ligature	Ligature S+P <sp></sp>	Graphemic	No ligature	l'esprit
Ligature	Ligature long S+L <fl></fl>	Graphemic	No ligature	meflent
Ligature	Ligature U+S <us></us>	Graphemic	No ligature	individus
Ligature	Ligature S+I di>	Graphemic	No ligature	quafi
Ligature	Ligature long S+long S (ff)	Graphemic	No ligature	ieuneffe
Ligature	Ligature F+F+I «ffi»	Graphemic	No ligature	fuffilance



Training and fine-tuning 1 Introduction

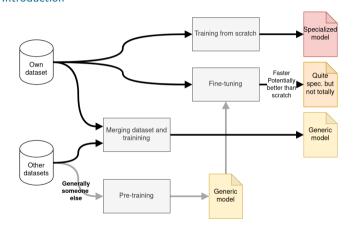


Figure: Example of workflows



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 Find and not always Accessible
 - uncertain formats & varying annotations
 - unclear **R**euse conditions



Introduction

► What is HTR-United?

▶ Why contribute?



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



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



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



- 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 schema 2 What is HTR-United?

⊙ 7 Open 🗸 8 Closed	Author +	Label 🗸	Projects +	Milestones +	Assignee +	Sort -
Add the ability to describe automatically aligned dataset #20 opened last week by Pontelneptique						
Add the ability to describe line pairs (image / text) #19 opened last week by Pontelseptique						
Let's start thinking about how to document models @vestion schema #16 opened on Dec 2, 2022 by alli-12						Ç 4
O Source manuscripts / images canonical links #15 opened on Nor 28, 2022 by PonteImplique						Ċ,
Add a "status" key enhancement properties (chema) #11 by allo: tz was closed on Jun 17, 2022						Ç 1
Add ORCID to Contributor objects (properties) software #10 by Ponteineptique was closed on Jun 17, 2022 \$\Phi \sigma_2 \$					6	Ç1 2
Transcription guidelines field: have a controled vocabulary of values properties as opened on Mar 22, 2022 by PonteIneptique	ema values					ÇJ 24
Add "German Kurrent" or "Kurrent" as a script wave #4 by thodel was closed on Jun 17, 2022 \$\phi_22 \$\phi_2 \$\phi_22 \$\phi_2					6	Ç) 8
Way around incomplete list of Scripts? (austion) schema (value) #1 by allx-tz was closed on Jun 17, 2022 (\$\phi \u0.2)					6	Ç) 6
© Contributor: have an Institution Name alternative to Name/Surname properties #2 by Postelneptique was closed on Jun 17, 2022 ♀ 402	schema					Фв
Schema: Add a "production software" key properties schema #3 by PonteIneptique was closed on Jun 17, 2022 \$\Psi vq2					5	C 5

Figure: Selection of issues open or closed in the GitHub repo dedicated to the schema



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)



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



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)