

"I'm here to fight for ground truth"

HTR-United, a solution towards a common for HTR training data

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

July 12 2023

DH2023, Graz, Austria





Situation with HTR and HTR Data

- 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 Reuse conditions

1 Introduction



Table of Contents

2 What is HTR-United?

▶ What is HTR-United?

- ► Why contribute?
- Conclusion



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

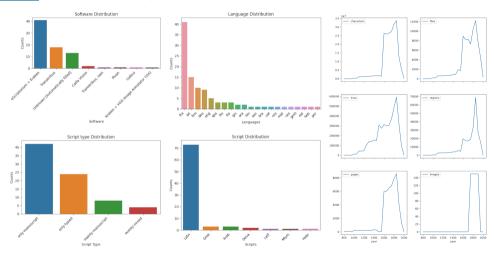


As of July 4th, 2023, the catalog refers to:

- 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



(Metrics overview) 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 schema 2 What is HTR-United?

⊙ 7 Open ✓ 8 Closed Author	→ Label →	Projects ▼	Milestones •	Assignee •	Sort ▼
O Add the ability to describe automatically aligned dataset #20 opened last week by Pontelneptique					
O Add the ability to describe line pairs (image / text) #19 opened last week by PonteIneptique					
O Let's start thinking about how to document models (question) (schema) #16 opened on Dec 2, 2022 by alise tz					Ç 4
O Source manuscripts / images canonical links #15 opened on Nov 28, 2022 by PonteIneptique					□ 3
O Add a "status" key enhancement properties schema #11 by allo: tz was closed on Ain 17, 2022					□ 3
Add ORCID to Contributor objects properties schema #10 by Ponteinepsique was closed on Am 17, 2022					□ 2
O Transcription guidelines field: have a controled vocabulary of values properties schema #5 opened on Mar 22, 2022 by Porteineptique					₩ 24
O Add "German Kurrent" or "Kurrent" as a script #4 by thodel was closed on Jun 17, 2022. \$\forall \tau \tau 0.2\$					□ s
② Way around incomplete list of Scripts? question schema (2005) #1 by alic-tz was closed on Jun 17, 2022. ¬ vo. 2 ¬ vo.					Q 6
○ Contributor: have an Institution Name alternative to Name/Surname properties (schema #2 by Postelreptique was closed on Jun 17, 2022 中 v02					□ 13
Schema: Add a "production software" key properties (schema) #3 by Postelneptique was closed on Jun 17, 2022 ↑ +0.2					Q 5

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



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)



Table of Contents

3 Why contribute?

- ▶ What is HTR-United?
- ► Why contribute?
- Conclusion



The stakes for the scientific community 3 Why contribute?

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



Table of Contents4 Conclusion

▶ What is HTR-United?

- ► Why contribute?
- **▶** Conclusion



- we need data papers with ground truth datasets
- kudos for datasets with a catalog entry and a data paper
- what about an HTR-United Data Journal?



Q & A

Thank you for listening! Your feedback will be highly appreciated!

PS: We present an example of reuse of data from HTR-United during the 2-3:30PM long paper session (LP-T3E) on Thursday!