“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 **F**ind and not always **A**ccessible
  - uncertain formats & varying annotations
  - unclear **R**euse conditions

Table of Contents

2 What is HTR-United?

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

It's a catalog

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

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

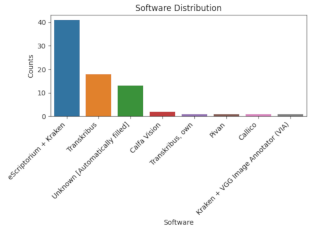
It's a catalog

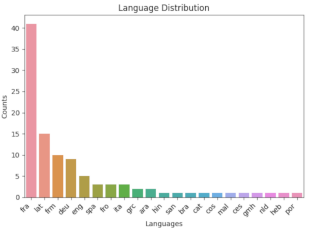
As of July 4th, 2023, the catalog refers to:

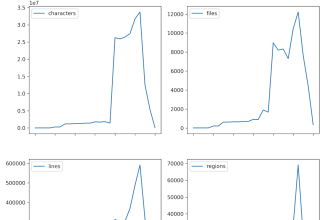
- 78 datasets created by at least 36 different projects
- 21 languages (a lot of French, and Latin) for 7 scripts (mostly Latin)
- handwritten and printed documents, mixed or "pure"
- a period going from 800 to 2023
- created with at least 6 different HTR software
- more than 44M characters, than 1M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

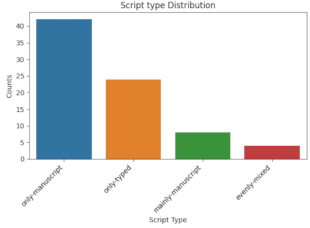
(Metrics overview)

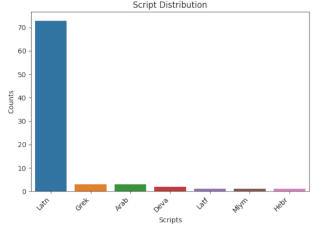
2 What is HTR-United?











It's a schema

- 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?

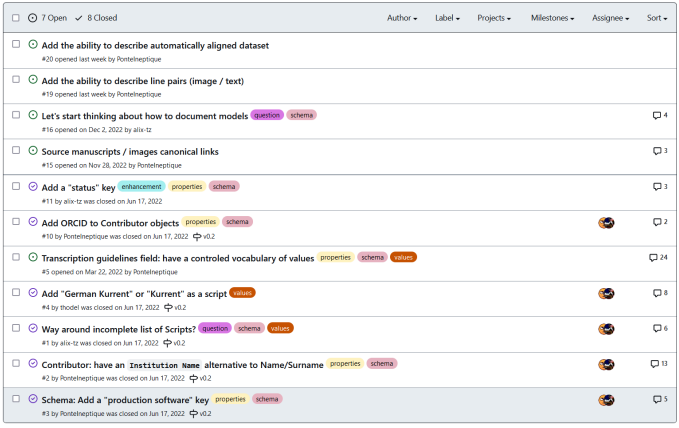


Figure: Selection of issues open or closed in the GitHub repo dedicated to the schema

It's a toolbox

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 Contents

4 Conclusion

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

The future

4 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!