



**HAL**  
open science

## Sharing HTR datasets with standardized metadata: the HTR-United initiative

Alix Chagué, Thibault Clérice

### ► To cite this version:

Alix Chagué, Thibault Clérice. Sharing HTR datasets with standardized metadata: the HTR-United initiative. Documents anciens et reconnaissance automatique des écritures manuscrites, CREM-MALab, Jun 2022, Paris, France. hal-03703989

**HAL Id: hal-03703989**

**<https://inria.hal.science/hal-03703989>**

Submitted on 24 Jun 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



# Sharing HTR datasets with standardized metadata: the HTR-United initiative

Alix Chagué (Inria, Université de Montréal)

Thibault Clérice (Ecole des chartes, Centre Jean Mabillon)



# The cost of data production

- OCR and HTR = fast and powerful tools for transcription
- collect text data; accelerate indexation of collections or build digital editions

but!

- we need for training examples (ground truth)
- it's expensive to produce and project don't always have the resources

so!

- it's helpful to rely of pre-existing models or pre-existing data
- especially to fine-tune models

# Findability of data

- it is difficult to find data and models
- we need greater habits of publishing them

Why it is hard to find?

- lack of clear keywords and locations to publish
- uncertainty on the formats
- uncertainty on the (copy)right status

Combining them is also difficult

- variation in annotations format
- variation in annotation rules

commun-au-htr

24 sept. 2020 à 4:33 PM

Et ça serait potentiellement intéressant de reprendre ce format de yml ( [github.com/cneud/ocr-gt/b...](https://github.com/cneud/ocr-gt/blob/master/.github/workflows/ocr-gt.yml) ) et d'avoir un github action ou équivalent pour updatater un [README.md](#) ?

(je te donne du travail d'une certaine manière...)

24 sept. 2020 à 4:34 PM

HTR United ?

24 sept. 2020 à 4:36 PM ✓

Et je peux t'aider à rédiger une description de l'idée

J'aime beaucoup HTR United



24 sept. 2020 à 4:36 PM

# This is why: HTR-United !

Merci pour cette petite collaboration express!  
J'espère que c'est le début de quelque chose de grand ! :)

24 sept. 2020 à 5:39 PM ✓

I

A catalog, a schema and some imperatives

# HTR-United's core: a catalog

```
- schema: https://htr-unique.github.io/schema/2022-04-15/schema.json
title: 'CREMMA-AN Testament De Poilus'
url: https://github.com/HTR-United/CREMMA-AN-TestamentDePoilus
description: "WWI\2019s Poilus' testaments edited by the Archives National
  \ the Testaments de Poilus project."
project-name: Testaments de Poilus
project-website: https://edition-testaments-de-poilus.huma-num.fr/
language:
- fra
script:
- iso: Latn
authors:
- name: Emmanuelle
  surname: de Champs
  roles:
  - project-manager
- surname: Clavaud
  name: Florence
  roles:
  - project-manager
  - quality-control
  - transcriber
- name: Pauline
  surname: Charbonnier
  roles:
```

Machine  
readable

Download the catalog

CREMMA-AN Testament De Poilus

Testaments de Poilus  
1914 - 1918

Link Data repository Link Citation File (CFF)

Language fra Script Latn Hands 1-per-file

Volume 33 652 characters 353 lines

Volume 105 regions

License CC-BY 4.0

Software eScriptorium + Kraken

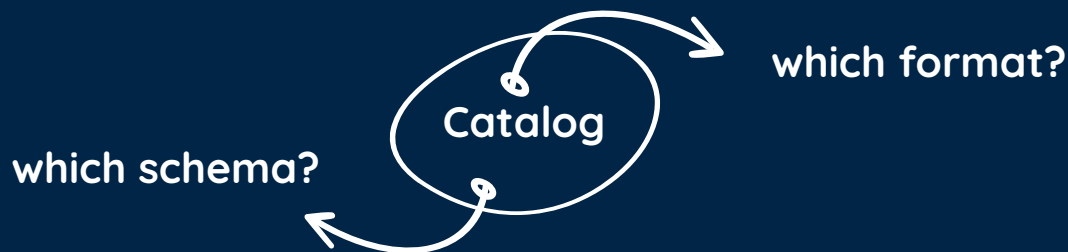
WWI's Poilus' testaments edited by the Archives National during the Testaments de Poilus project.

Human  
Readable

Browse the Catalog

- provide access to the datasets!
- foster findability (filtering)
- shed light on the importance of citation

# The schema: technical choices



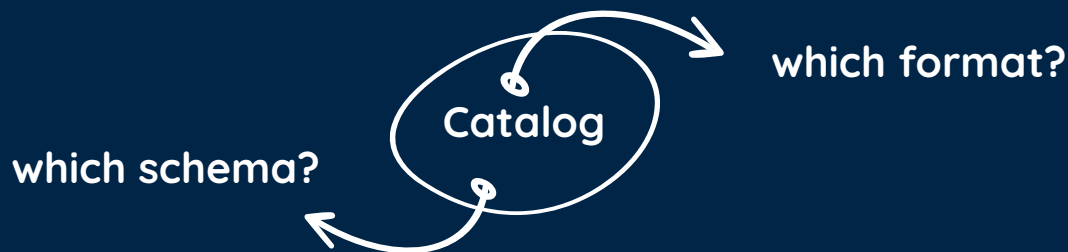
- no pre-existing shared way to describe transcription ground truth
- writing and controlling a entry in the catalog had to stay simple/easy

3 formats options :

- XML
- JSON
- YAML



# The schema: technical choices

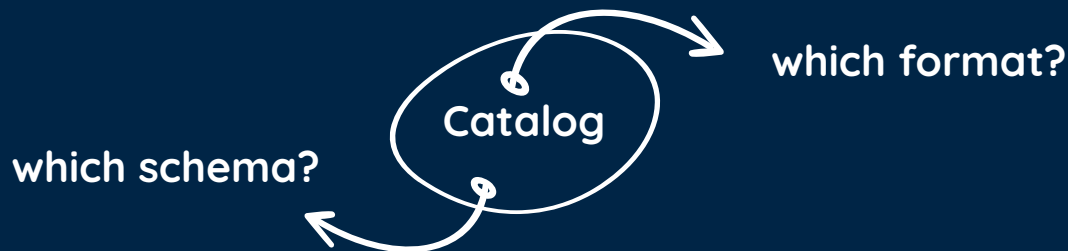


- no pre-existing shared way to describe transcription ground truth
- writing and controlling a entry in the catalog had to stay simple/easy

3 formats options :

- XML
- JSON
- YAML + JsonSchema

# The schema: technical choices



- no pre-existing shared way to describe transcription ground truth
- writing and controlling a entry in the catalog had to stay simple/easy

3 formats options :

- XML
- JSON
- YAML + JsonSchema

Why a schema?

- a machine actionable way to check conformity of a submission to the catalog
- clearer documentation on the evolutions of the schema

# The schema: what does it include

- dataset title, link and short description
- license of use and format used for data (ALTO vs PAGE)
- software used to produce the data
- period covered by the documents
- languages and the scripts
- collaborators of the project
- script type (e.g. print vs. manuscript or mixed)
- number of hands or of fonts
- metrics (mainly lines, characters and files)

## Controlled vocabularies:

- format
- language
- script
- script type
- number of hands/fonts
- collaborators' roles

# The schema: about transcriptions

- currently: 1 field of free text for “transcription guidelines”

The longer term goals:

- automatic parsing of the ground truth (character sets, and normalization systems)
- build a controlled vocabulary for transcription guidelines

examples:

Transcription guidelines

The original transcriptions were performed on a crowdsourcing application (<https://testaments-de-poilus.huma-num.fr/#!/>) under the supervision of the Archives nationales de France. Only the allographic portions of the documents were transcribed. Any marginal elements added later by clerks or archivists are neither segmented nor transcribed. The segmentation follows the SegmOnto ontology. Abbreviations and misspelling were not corrected. Superscripted portions of text are preceded by ^.

Transcription guidelines

See: <https://www.koenigsfelden.uzh.ch/exist/apps/ssrq/intro.html#richtlinien>

Transcription guidelines

Diplomatique, mais pas allographétique.

Transcription guidelines

Kept abbreviation and transcribed long s as long s

## Transcription guidelines field: have a controlled vocabulary of values #5

Open Pontelneptique opened this issue on 22 Mar · 23 comments



Pontelneptique commented on 22 Mar

Member

This one has been in my head for quite a long time.

Right now, we have free text, which means it is not machine actionable. I'd like to have the ability to populate a list of acceptable value, such as Resolved Abbreviation, Unresolved Abbreviation, Corrected Spelling, Original Spelling, Special character used (eg. MUFI), things like that. It would go alongside the transcription guidelines but would make the whole thing a little more machine actionable.

Pontelneptique added schema values properties labels on 22 Mar

# Imperatives


- a collaborative enterprise for the community
- friendly to consumer and producers of data
- as low tech as possible (because \$\$)




# Imperatives: Welcoming to users

- we are editors of the catalog
- but we have our own domains of expertise and interest (mainly French from the medieval period to today)
- we don't know every use case
- therefore the initiative benefits from users' contributions to the ecosystem

Add "German Kurrent" or "Kurrent" as a script #4

 Closed thodel opened this issue on 19 Mar · 8 comments

 thodel commented on 19 Mar

Add "German Kurrent" or "Kurrent" as a script-type in your form.  
Since kurrent is not to be combined with latin scripts, a differentiation makes sense.

We added a qualifier option for each item in to the ISO list of scripts

# Imperatives: User friendly

Even if YAML is easy, it is not *that* easy

- we created a form to help generate the catalog entry
- updating the form along with the schema takes time but it's worth it
- we want a form as simple as possible, even though the schema gains complexity

We use Continuous Integration via Github Actions. For each update in the catalog:

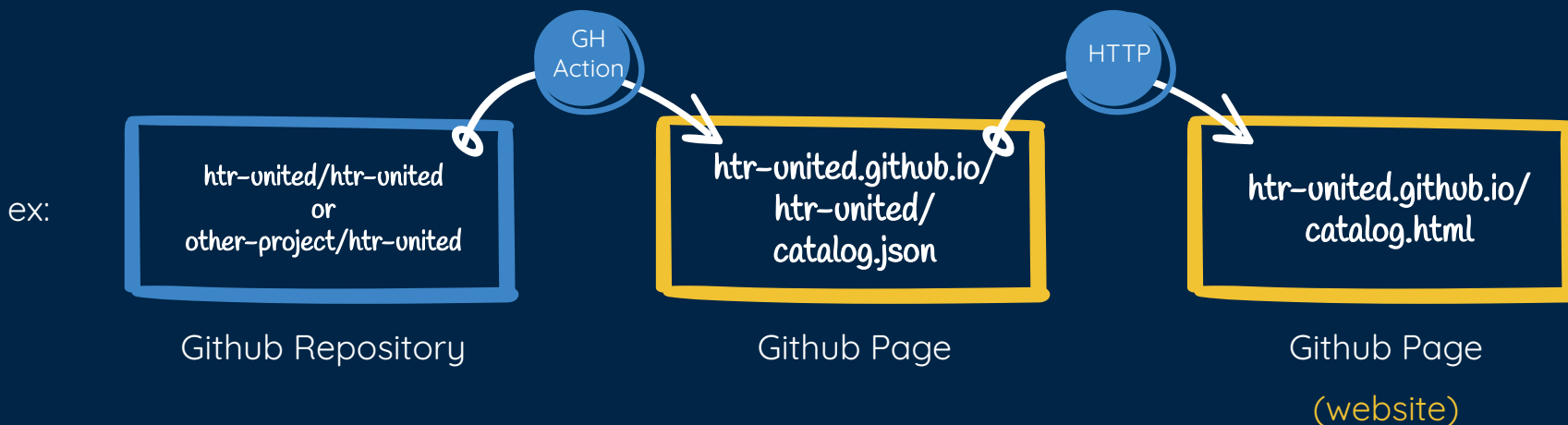
- automatic validity control of the metadata (.yml)
- automatic validity control of the files (.xml)
- automatic generation of metrics (.yml and badges)

CI requires skills, for we made a form for that too!



# Imperatives: Low Tech

- we are not funded but we have ambitions
- **Github Pages** support **HTML + CSS + JavaScript**
- **Github Actions** can be used to build **JSON** files fed to the website





II

Ecosystem

# Ecosystem: Consolidating workflows

- we started with cataloging and normalizing but didn't (couldn't ?) stop there
- four common actions between different projects:
  - an htr-united.yml catalog entry
  - people want metrics on their dataset
  - people want cohesion within their dataset
  - people want compatibility between their datasets

We developed or integrated tools to address these actions

# Ecosystem: Control the catalog with HTRuc

Why?

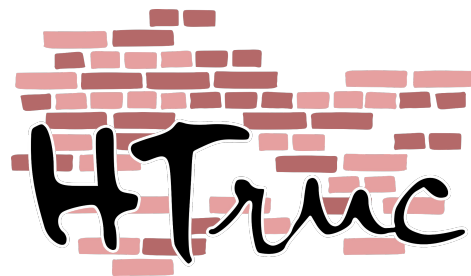
- users provide catalog entries as YAML files, which content is added to the core catalog
- we need to control this entry before adding it to the catalog

What is it?

- a command line tool to control the validity of the htr-united.yml files

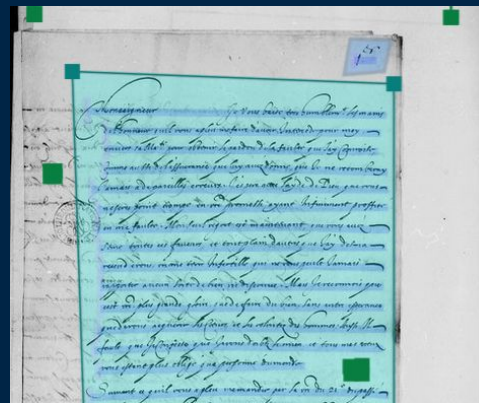
Other features include:

- automatic conversion to the next version of the schema
- update record metrics
- collect records across different folders or Github repositories
- aggregating the records in the core catalog



# Ecosystem: Control the data with HTRVX

- control schema validity in XML files
- can control conformity to ontologies like SegmOnto (correctness of the tagging)
- can control occurrences of empty elements (lines or regions)



## HTRVX

failed on 9 May in 1m 52s

Search logs

```
199 ... Testing data/manuscrit-verne-20000/f13.xml
200 ✓ Segmonto test for data/manuscrit-verne-20000/f13.xml: Valid
201 x Detection of empty lines or region in data/manuscrit-verne-20000/f13.xml: Empty elements founds (1)
202   → 1 empty zone(s) found: #eSc_textblock_6d136cdc
203   → 0 empty line(s) found:
204 ✓ Schema for data/manuscrit-verne-20000/f13.xml: Valid
205 ... Testing data/manuscrit-verne-20000/f170.xml
206 ✓ Segmonto test for data/manuscrit-verne-20000/f170.xml: Valid
207 x Detection of empty lines or region in data/manuscrit-verne-20000/f170.xml: Empty elements founds (1)
208   → 1 empty zone(s) found: #eSc_textblock_86ce7418
209   → 0 empty line(s) found:
210 ✓ Schema for data/manuscrit-verne-20000/f170.xml: Valid
```



# Ecosystem: Update information with HUMG(enerator)

Why?

- we like metrics, you like metrics
- but computing metrics takes time and is boring

What is it?

- a command line tools actionable with Github Actions
- creates badges
- update the YML file with the metrics

## Notaires de Paris - Répertoires

Ground truth for various Parisian notary's registers (centuries)

license CC-BY DOI 10.5072/zenodo.977691

XML Files 218 Regions 1138 Lines 29410 Characters 525018

## CREMMA Early Modern Books

Characters 84726 Regions 451 Lines 2603 XML Files 98



# Ecosystem: Control the transcription with ChocoMufin

- originally developed for the CREMMA medieval corpus (Pinche & Clérice 2021)
- originally tracked variations of encoding for medieval characters

Why?

- consistency in character sets is sometimes hard to maintain

What does it do?

- control character or sequences of characters in a dataset
- convert characters according to a conversion table



III

Perspectives by way of concluding

# Perspectives: A dataset builder

- browsing and reading the description requires a human for the free text sections
- we want to add more filters to add automation
- we could have a lab bench to combine and compose new datasets



# Perspectives: Control vocabularies in the schema?

- necessary to add automation over fields like “transcription guidelines”

## Transcription guidelines field: have a controlled vocabulary of values #5

 Open Pontelneptique opened this issue on 22 Mar · 23 comments



Pontelneptique commented on 22 Mar

Member  

This one has been in my head for quite a long time.

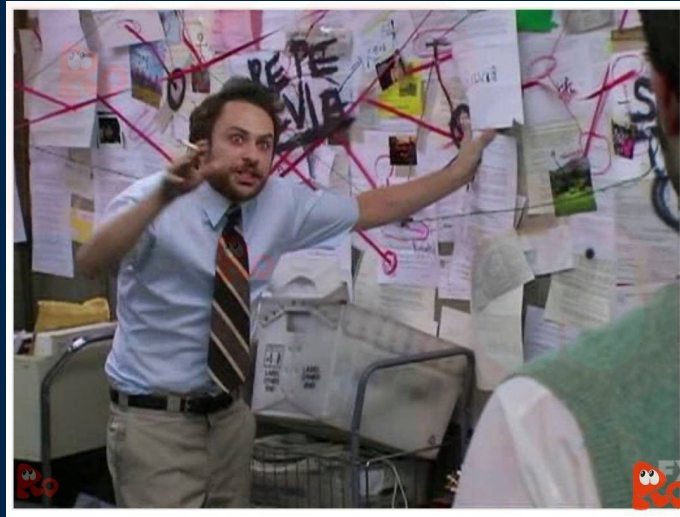
Right now, we have free text, which means it is not machine actionable. I'd like to have the ability to populate a list of acceptable value, such as `Resolved Abbreviation`, `Unresolved Abbreviation`, `Corrected Spelling`, `Original Spelling`, `Special Character used` (eg. MUF1), things like that. It would go alongside the transcription guidelines but would make the whole thing a little more machine actionable.



Pontelneptique added `schema` `values` `properties` labels on 22 Mar

# Perspectives: Tools as web services?

- offer server versions of the tools
- alternatives to running them locally or via Github Actions
- but requires having servers





# Perspectives: Provide generic models

- An out-of-the-box model for **Modern and contemporary manuscripts in French**
- **1600 to 2022**, includes 100 pages of 19th Spanish and 20th English and some prints
- trained on 792 files (**72 237 lines**) from **13 datasets** (+1 private)
- Accuracy: **90,6% on devset**, 79.2% on 21th century testset (CREMMAWiki), **81,5% on Quicherat** (eval with KaMI)



Croix du jubé à S. Pierre d.e Louvain.  
Elle est du &XV<sup>e</sup> sHièoclae, en bois, fleuronnée s<s>  
<uo> de quatref ifceuille,r et de fleurs de lepys et aengraeélé  
Ssur dsaes contours. Le tout, porte dsutre un picoeid  
qui lui-même reposea siur un petit rochaer et  
cuenne- grand.e travers-e paur dessous, d'ouñ  
Ss'élève à chaque bout une figuriee ,; d'iun côté  
la vierge et de l'auctre St. Jean. Sours la  
maême traverse,q un tableau à  
troise rvolets, avec deux piedtts qui  
descendent pooceusr faire pa<r> l'aoffice de  
poeintes d'appui sutr le pjiubé.  
Les feuillures du tableau deu  
côtié de la nef tsont occeunpgeés  
par trois figures Ssculpteúes.  
Dans les fleurons de la crsoifx  
Fson les arnimaux symhboleiques.

	Default	Ignoring digits	Ignoring case	Ignoring punctuation	Ignoring diacritics	Combining all options
Levensthein Distance (Char.)	115	115	112	108	90	80
Levensthein Distance (Words)	62	62	61	59	55	50
Word Error Rate (WER in %)	53.448	53.448	52.586	52.678	47.413	44.642
Char. Error Rate (CER in %)	18.459	18.459	17.977	18.12	14.446	13.422
Word Accuracy (Wacc in %)	46.551	46.551	47.413	47.321	52.586	55.357

Thank you! 🙏

- <https://htr-united.github.io/>
- <https://github.com/HTR-United/htr-united>
- <https://github.com/HTR-United/schema>

Don't hesitate to start a conversation about use cases in an issue!