

"I'm here to fight for ground truth": HTR-United, a solution towards a common for HTR training data

Alix Chagué, Thibault Clérice

▶ To cite this version:

Alix Chagué, Thibault Clérice. "I'm here to fight for ground truth": HTR-United, a solution towards a common for HTR training data. Digital Humanities 2023: Collaboration as Opportunity, Alliance of Digital Humanities Organizations; University of Graz, Jul 2023, Graz, Austria. hal-04094233

HAL Id: hal-04094233 https://inria.hal.science/hal-04094233

Submitted on 7 Jun2023

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 - ShareAlike 4.0 International License

Proposal to DH2023: Collaboration as Opportunity "I'm here to fight for ground truth": HTR-United, a solution towards a common for HTR training data

Alix Chagué^{1,2,3} and Thibault Clérice^{1,4}

 ¹ALMAnaCH - Automatic Language Modelling and Analysis & Computational Humanities, Inria, Paris, France
 ²UdeM - Université de Montréal, Montréal, Canada
 ³EPHE - École Pratique des Hautes Études, Paris, France
 ⁴CJM - Centre Jean Mabillon, Paris, France

November 2022

The growth of computation power and rise of artificial intelligence (in particular Deep Learning) allowed for the development of automatic text recognition, both on printed texts (OCR) and handwritten ones (HTR). Such technologies can now make millions of images of texts from various periods of time, held in patrimonial institutions, available for further search and processing.

HTR became more accessible when user friendly interfaces started to be developed: namely Transkribus from 2015 [6] and eScriptorium from 2019[5]. In the case of HTR and old prints though, one of the hurdles remaining to be overcome is the access to robust models capable of recognizing coherent texts in spite of the multiple variations in handwriting or fonts. Such models usually necessitate users to produce large amounts of manual transcriptions considered as perfect -called ground truth-, which is a costly task. It requires a good understanding of the way deep learning functions, skills in paleography, and time. An easy way to reduce the costs of creating the training data to obtain a model is to rely on the data produced by other projects. Unfortunately, they are hard to find and not always published, because there is no incentive to put in this extra effort, neither for their publication nor for their documentation.

HTR-United is a collaborative initiative who's main purpose is to improve the findability of these open datasets, covering as many periods, scripts and languages as possible. Through this initiative, we support the creation a public catalog of dataset descriptions, contributed by individuals volunteering their own datasets. In general, descriptions are submitted as a YAML file filled with the help of a form available

es champs dont le nom est suivi d'un * sont oblig	atoires.	
General Information abou	t the dataset	
repository's or dataset's title"	CREMINA Manuscrits ou 19e https://github.rom/HTR-Inited/CREMMA-MSS-19	
ihort description*	Examples of handwritten texts from various authors from the 19th century.	
ink to a CITATION.cff file	Link to a CITATION.df file	
	CIAIUNUT tes allow for generating quick are obtained or baller ones on pithub for example. See <u>integritation-testament aptivulue</u>	
icense"	CC-BY 4.0 Format standard* ALIO JONE	
oftware use to produce the data*	eScriptorium + Kraken	
	You can use the following buttons to escriptorium - Kraken Transkribus OCR For All populate the field.	
General information abou	It the project	
troject's name (if different from the epository/dataset's one)	CREMMA	
ink to project's website	Link to project's website (if applicable)	
uthority and role(s)	Add a member	
	Thibault Clérice	
	ORCID 0000-0003-1852-9204	
	Contributor is an institution Roles Control Contro Control Control Control Contro	ntrol
	□ Alignment □ Digitizatio ✔ Project Manager □ Support	n
	Add a member Remove a member	
	(\mathbf{a})	
	(a)	
Get formatted	metadata	
Gertomateo		
scheme: https://htr-united.github.io/	schemø/2022-04-15/schemø.json	
<pre>title: CREMMA Manuscrits du 19e url: https://github.com/HTR-United/CR authors:</pre>	EMMA-MSS-19	
- name: Thibault surname: Clérice		
orcid: 0000-0003-1852-9284 roles:		
 project-manager quality-control 		
- name: Alix surname: Chagué		
orcid: 0000-0002-0136-4434 roles:		
 project-manager quality-control 		
- name: Baudoin surname: Davoury		
- transcriber - aligner		
 name: Soline surname: Doat 		
roles: - transcriber		
- aligner - name: Margaux		
surname: Faure orcid: 0000-0001-5815-9506		
- transcriber		
- name: Maxime surname: Humeau		
roles: - transcriber		
- aligner institutions: []		
description: Examples of handwritten project-name: CREMMA	texts from various authors from the 19th century.	
 Inguage: fra production coftware: efections + K 	raian	
script:	rancei	
	(h)	
	(0)	
HTR-United / htr-united	ublic)	
↔ Code ⊙ Issues 13 I'l Pull	requests 😡 Discussions 💿 Actions 🖽 Projects 🖽 Wiki	
Adding dataset CREM	IMA-MSS-19	
Write Preview		
Hello I I would like to sub	mit the description of a new dataset for the HTR-United catalog	
Here is our dataset YAML	file:	
schema: https://htr- title: CREMMA Manusc	united.github.io/schema/2022-04-15/schema.json rrits du 19e	
url: https://github. authors:	com/HTR-United/CREMMA-MSS-19	
 name: Thibault surname: Clérice 		
roles:	1852-9284	
- project-manag - quality-contr	ier rol	
- name: Alix surname: Chagué	0126-0020	
roles:	0130-4434	
- quality-contr - name: Baudoin	01	
surname: Davoury roles:		
- transcriber - aligner		
- name: Soline surname: Doat		
roles: - transcriber		
- aligner - name: Margaux		
surname: Faure orcid: 0000-0001-	5815-9506	
- transcriber		
- name: Maxime surname: Humeau		
roles:		
	(C)	

Figure 1: (a) Excerpt of the form to record the description of a new dataset ; (b) YAML content generated by the form ; (c) YAML description of a dataset submitted to HTR-United with Github. 2

R-United	Q						
CREMMA Manuscrits du 19e		CREMMA	Filters				
1800 - 1899			Language	Language All languages			
Language fra Script Latn	Script Type 🖌 only-manuscript Hands 1-per-folder		Script	All languages			
S Volume 55,581 characters S Volume 69 files S Volume 1,807 lines S Volume 167 regions			Script type	All			
License CC-8Y 4.0			Project	All projects			
Software eScriptorium + Kraken			Dates	Not before:	I Not after	a 10	
Manuscripts of the 19th century		Description	Number of results	58			
Authors: Clérice, Thibault and Cha and Humeau, Maxime	gué, Alix and Davoury, Baudouin and Doat, Soline	e and Faure, Margaux	Statistics about	Unit	Amount	Projects with this uni	
		Transarintian guidalines	the results	Characters	42,840,534	51	
Abréviations conservées.		Transcription guidennes		Lines	771,012	5	
Bibtex 🏢		Citation	Options	Show Transcription Guidelines			
@misc{htr_united_https-gith type = dataset,	ubcom-htr-united-cremma-mss-19,				on mormations		

(a) Record's view.

(b) Catalog filter view.

Figure 2: View of records in the catalog: records can be see in their own page (a) or browsed in the catalog, including after using filters (b).

on HTR-United website (Figure 1)¹. Raising awareness on the necessity to correctly document such shared datasets, HTR-United favors the implementation of the FAIR principles in the specific case of text recognition training datasets[2]. The catalog² can be browsed using filters (script, language, type of font, period, etc.) and offer means to easily cite a dataset (Figure 2).

The initiative is set up as an ecosystem of public Github repositories³, which guarantees the existence of precious versioning features for an ever-evolving catalog, transparency from all the parties as well as the possibility for us to rely on minimalistic developments. For example, anytime a dataset description is validated by our team, a Github Action processes all the existing descriptions in order to generate a new version of the catalog in the form of a pivot YAML file⁴: the catalog is never directly edited manually which reduces the risks of introducing errors. While a repository is dedicated to gathering all the descriptions feeding the catalog, another one hosts the specifications of the schema used to control the conformity of the descriptions⁵. Anyone can open a discussion to suggest the addition of new features in the specifications, or access the details of the arguments having led to the modification of the schema. Additionally, we aim to provide and maintain a suite of tools, available locally or through Github Actions and continuous integration, which help control, document and manage dataset on the short and long term, specifically in heavily collaborative contexts⁶.

During the DH2023 conference, we would like to introduce HTR-United to the

 $^{^{1}\}mathrm{See}$ https://htr-united.github.io/document-your-data.html.

 $^{^{2}\}mathrm{See} \ \mathtt{https://htr-united.github.io/catalog.html}.$

³See https://github.com/HTR-United.

⁴See in particular https://github.com/HTR-United/htr-united/blob/master/htr-united.yml.

⁵See https://github.com/HTR-United/schema

⁶See https://htr-united.github.io/actions.html

international DH community by presenting how the ecosystem is organized, how contributors can submit new entries to the catalog as well as the stakes of contributing to such an initiative. HTR-United can be useful for the entire community of users of HTR technologies as the datasets listed in the catalog cover more and more languages or writing systems.

We would like to present some of the most interesting outcomes of such a collaborative catalog. First, various generic models for HTR were trained thanks to having access to a great variety of ground truth datasets ⁷, which are of tremendous importance for the successful development of HTR for the humanities and the cultural institutions. The existence of such models allows smaller institutions or groups of researchers to quickly train robust models by simply fine-tuning generic models in stead of starting from scratch [3]. Secondly, one of the most exciting aspects of possessing such a space for exchanging information about ground truth datasets is the fact that it creates opportunities to pave the way towards a (international) standardization of transcription practices in the context of ground truth creation.

References

- Alix Chagué and Thibault Clérice. HTR-United Manu McFrench V1 (Manuscripts of Modern and Contemporaneous French). Version 1.0.0. June 2022. DOI: 10. 5281/zenodo.6657809. URL: https://doi.org/10.5281/zenodo.6657809.
- [2] Alix Chagué and Thibault Clérice. "Sharing HTR datasets with standardized metadata: the HTR-United initiative". In: *Documents anciens et reconnaissance automatique des écritures manuscrites*. The recording of the conference is available at: https://www.canal-u.tv/chaines/enc/25-sharing-htr-datasets-with-standardizedmetadata-the-htr-united-initiative. CREMMALab. Paris, France, June 2022. URL: https://hal.inria.fr/hal-03703989.
- [3] Alix Chagué, Thibault Clérice, and Laurent Romary. "HTR-United : Mutualisons la vérité de terrain !" In: DHNord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux. MESHS. Lille, France, Nov. 2021. URL: https://hal.archives-ouvertes.fr/hal-03398740.
- [4] Tobias Mathias Hodel et al. "General models for handwritten text recognition: feasibility and state-of-the art. German kurrent as an example". In: *Journal of open humanities data* 7.13 (2021), pp. 1–10.
- [5] Benjamin Kiessling et al. "eScriptorium: an open source platform for historical document analysis". In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). Vol. 2. IEEE. 2019, pp. 19–19.

⁷See the CREMMA Medieval model[7], the Manu McFrench models [1] and other experiments on large training datasets such as Hodel et al. [4].

- [6] Guenter Muehlberger et al. "Transforming scholarship in the archives through handwritten text recognition". In: Journal of Documentation 75.5 (Jan. 2019), pp. 954–976. ISSN: 0022-0418. DOI: 10.1108/JD-07-2018-0114. URL: https: //doi.org/10.1108/JD-07-2018-0114.
- [7] Ariane Pinche. "Generic HTR Models for Medieval Manuscripts The CREM-MALab Project". working paper or preprint. Nov. 2022. URL: https://hal. archives-ouvertes.fr/hal-03837519.