



HAL
open science

Workshop HTR-United: metadata, quality control and sharing process for HTR training data

Thibault Clérice, Alix Chagué, Hugo Scheithauer

► To cite this version:

Thibault Clérice, Alix Chagué, Hugo Scheithauer. Workshop HTR-United: metadata, quality control and sharing process for HTR training data. DH 2023 - Digital Humanities Conference: Collaboration as Opportunity, Alliance of Digital Humanities Organizations; University of Graz, Jul 2023, Graz, Austria. hal-04094235

HAL Id: hal-04094235

<https://inria.hal.science/hal-04094235v1>

Submitted on 7 Jun 2023

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

Proposal to DH2023: Collaboration as Opportunity
Workshop HTR-United: metadata, quality control
and sharing process for HTR training data

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

¹ALMAnaCH - Automatic Language Modelling and Analysis &
Computational Humanities, Inria, Paris, France

²CJM - Centre Jean Mabillon, Paris, France

³UdeM - Université de Montréal, Montréal, Canada

⁴EPHE - École Pratique des Hautes Études, 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 [3] and eScriptorium from 2019[2]. 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-, taking the form of pairs of images and transcriptions (XML files containing the coordinates and the corresponding text), 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 whose 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 of 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

Les champs dont le nom est suivi d'un * sont obligatoires.

General information about the dataset

Repository's or dataset's title* CREMMA Manuscrits du 19e

Link to repository* <https://github.com/HTR-United/CREMMA-MSS-19>

Short description* Examples of handwritten texts from various authors from the 19th century.

Link to a CITATION.cff file [Link to a CITATION.cff file](#)
CITATION.cff files allow for generating quick API citations or BibTeX ones on github for example. See <https://citation-file-format.github.io/>

License* CC-BY 4.0 Format Standard* ALTO XML

This is a non-exhaustive selection of license options.

Software used to produce the data* eScriptorium - Kraken
You can use the following buttons to populate the field: eScriptorium - Kraken Transcriber OCR For All

General information about the project

Project's name (if different from the repository/dataset's one) CREMMA

Link to project's website [Link to project's website \(if applicable\)](#)

Authority and role(s)

Add a member

Thibault Clérice

ORCID: 0000-0003-1852-9204

This contributor is an institution

Roles: Transcription Alignment Project Manager Quality Control Digitization Support

Add a member Remove a member

(a)

Get formatted metadata

```


schema: https://htr-united.github.io/schema/2022-04-15/schema.json
title: CREMMA Manuscrits du 19e
url: https://github.com/HTR-United/CREMMA-MSS-19
authors:
  - name: Thibault
    surname: Clérice
    orcid: 0000-0003-1852-9204
    roles:
      - project-manager
      - quality-control
  - name: Alix
    surname: Chagud
    orcid: 0000-0002-0136-4434
    roles:
      - project-manager
      - quality-control
  - name: Baudouin
    surname: Davoury
    roles:
      - transcriber
  - name: Soline
    surname: Doat
    roles:
      - transcriber
  - name: Margaux
    surname: Faure
    orcid: 0000-0001-5015-9006
    roles:
      - transcriber
  - name: Maxime
    surname: Mumeau
    roles:
      - transcriber
  - name: Maxime
    surname: Mumeau
    roles:
      - aligner
Institutions: []
description: Examples of handwritten texts from various authors from the 19th century.
project-name: CREMMA
language:
  - fra
production-software: eScriptorium - Kraken
script:

```

(b)

HTR-United / htr-united Public

<> Code Issues 13 Pull requests Discussions Actions Projects Wiki

 Adding dataset CREMMA-MSS-19

Write Preview

Hello ! I would like to submit the description of a new dataset for the HTR-United catalog!

Here is our dataset YAML file:

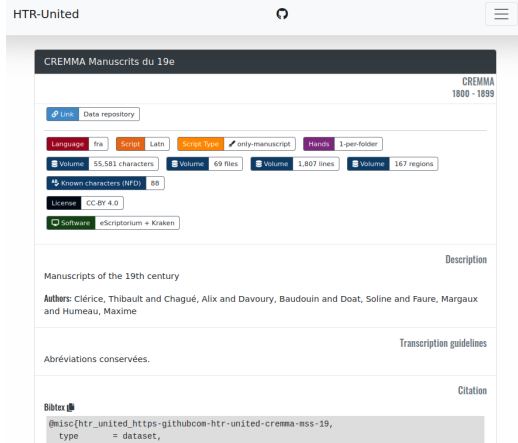
```

schema: https://htr-united.github.io/schema/2022-04-15/schema.json
title: CREMMA Manuscrits du 19e
url: https://github.com/HTR-United/CREMMA-MSS-19
authors:
  - name: Thibault
    surname: Clérice
    orcid: 0000-0003-1852-9204
    roles:
      - project-manager
      - quality-control
  - name: Alix
    surname: Chagud
    orcid: 0000-0002-0136-4434
    roles:
      - project-manager
      - quality-control
  - name: Baudouin
    surname: Davoury
    roles:
      - transcriber
  - name: Soline
    surname: Doat
    roles:
      - transcriber
  - name: Margaux
    surname: Faure
    orcid: 0000-0001-5015-9006
    roles:
      - transcriber
  - name: Maxime
    surname: Mumeau
    roles:
      - aligner
Institutions: []
description: Examples of handwritten texts from various authors from the 19th century.
project-name: CREMMA
language:
  - fra
production-software: eScriptorium - Kraken
script:

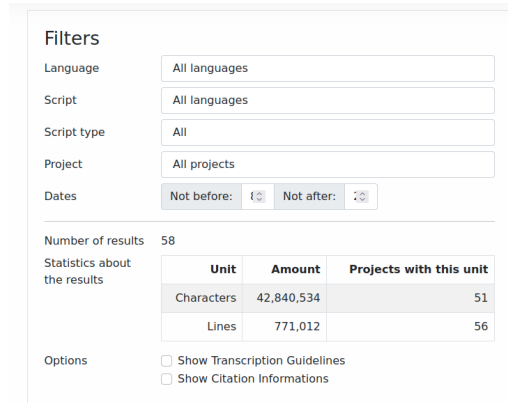
```

(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.



(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 principles² in the specific case of text recognition training datasets[1]. The catalog³ 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 repositories⁴, 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 minimalist 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 file⁵: 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 descriptions⁶. 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 contexts⁷.

¹See <https://htr-unique.github.io/document-your-data.html>.

²The letters stand for Findable, Accessible, Interoperable and Reusable. For more information, see <https://www.go-fair.org/fair-principles/>

³See <https://htr-unique.github.io/catalog.html>.

⁴See <https://github.com/HTR-United>.

⁵See in particular <https://github.com/HTR-United/htr-unique/blob/master/htr-unique.yml>.

⁶See <https://github.com/HTR-United/schema>

⁷See <https://htr-unique.github.io/actions.html>

During the DH2023 conference, we would like to organize a workshop focused on three essential aspects of publishing ground truth: 1) the architecture of such a dataset, 2) its description to make it findable and reusable by third-party users, and 3) mechanisms for longer term quality control.

The workshop will take place during a 4 hour long session (half a day). After briefly presenting the context of creation of HTR-United and its overall architecture, we will first examine our template for building ground truth repositories⁸. This template is useful to highlight the essential elements which must be found in such a dataset: the transcriptions and images (or links to images), information about the context of production and about the source document(s), a license, etc. The second phase of the workshop will focus on how to create the description of a ground truth dataset in order to add it to HTR-United using the aforementioned form and how to submit the resulting catalog entry. We hope that this stage will be the occasion to longer discuss the choices made during the construction of the metadata schema and potential ways to improve the existing standards. Lastly, we will introduce the suite of tools designed to help manage and control the content of the repositories and/or its description in HTR-United. This suite includes HUMGenerator (for the generation of additional metadata), HTRVX (to control the validity of the XML files containing the ground truth), and ChocoMufin (which controls the list of characters used in a dataset)⁹. We will demonstrate how they can be used locally as well as through Github Actions (for datasets hosted on Github).

The targeted audience would benefit from being familiar with the basis of hand-written text recognition processes as well as with environments such as Github. However, no technical skill is required since HTR-United and its suite of tools does not require any local installation. Attendee possessing datasets of ground truth for HTR will be welcome to use their own dataset as examples during the workshop.

After this workshop, an attendee will:

1. Be able to use HTR-United's template to create a properly structured and documented dataset of ground truth for HTR;
2. Know how to use HTR-United's form and catalog to submit a dataset description or find datasets useful to their project;
3. Know how to apply HTR-United suite of tools to control the quality of the ground truth in the dataset and generate up-to-date metadata; and
4. Be further acquainted with the notion of continuous integration which can be useful in many contexts, way beyond the scope of HTR technologies.

⁸See <https://github.com/HTR-United/template-htr-united-datarepo>

⁹For more details about these tools, see <https://htr-united.github.io/tools.html>.

1 Instructors

1.1 Thibault Clérico

Thibault Clérico is a digital humanist with a classical studies background, who served as an engineer both at the Centre for eResearch (Kings College London, UK) and the Humboldt Chair for Digital Humanities (Leipzig, Germany) where he developed the data backbone of the future Perseus 5 (under the CapiTainS.org project). He was the head of the DH applied to GLAM program for 5 years at the École nationale des Chartes. He is a founding member of the Technical Committee for the Distributed Text Services standard (w3id.org/dts), and co-founder of HTR-United. The major part of his teaching is dedicated to cultural heritage data engineering, development good practices, standards for communication and programming languages. His research mainly focus on natural language processing for ancient languages through deep learning, the distribution of corpora and computational methods applied to the humanities.

1.2 Alix Chagué

Alix Chagué is a PhD student in Digital Humanities affiliated to the ALMAAnaCH team at Inria (Paris, France) and the CRIHN (*Centre de Recherche Interuniversitaire sur les Humanités Numériques*) at the University of Montreal (Montreal, Canada). Her research interest are focused on the development of clearer methodologies to apply automatic transcription techniques (such as HTR) by patrimonial institutions and researchers in the DH community. She co-founded HTR-United and, as a Research and Development engineer from 2018 to 2021, she contributed to various projects involving the automatic recognition of handwritten texts: ANR TIMEUS, LECTAUREP, and eScriptorium.

References

- [1] Alix Chagué and Thibault Clérico. “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-standardized-metadata-the-htr-united-initiative>. CREMMALab. Paris, France, June 2022. URL: <https://hal.inria.fr/hal-03703989>.
- [2] 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.
- [3] 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>.