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

Clérice, Thibault

clerice.thibault@algorythme.net
Centre Jean Mabillon, PSL-Ecole nationales des chartes; ALMAnaCH, Inria, France

Chagué, Alix

alix.chague@inria.fr
ALMAnaCH, Inria, France; Université de Montréal, Montréal, Canada

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 (Muehlberger et al. 2019) and eScriptorium from 2019 (Kiessling et al. 2019). 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 despite 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 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 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 [2] in the specific case of text recognition training datasets (Chagué / Clérice 2022b). The catalog [3] can be browsed using filters (script, language, type of font, period, etc.) and offer means to easily cite a dataset (Figure 2).

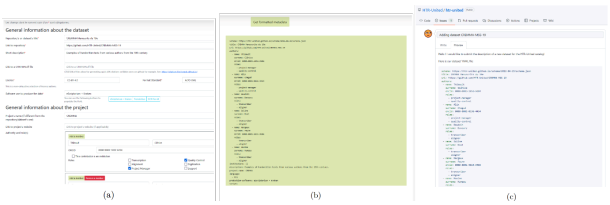
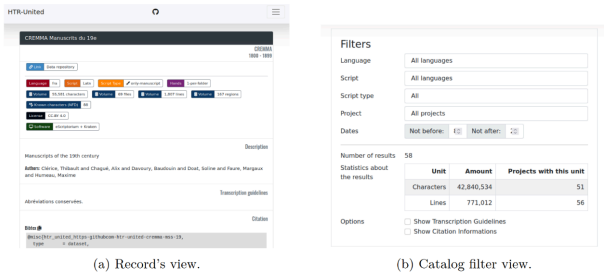


Figure 1: (a) Excerpt of the form to record the description of a new dataset; (b) YAML content
generated by the form; (c) YAML description of a dataset submitted to HTR-United with Github.



(a) Record's view. (b) Catalog filter view.

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

The initiative is set up as an ecosystem of public Github repositories [4], 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 file [5]: 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 descriptions [6]. 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 contexts [7].

During the DH2023 conference, we will 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 repositories [8]. 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 handwritten 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.

Instructors

Thibault Clérice

Thibault Clérice 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.

Alix Chagué

Alix Chagué is a PhD student in Digital Humanities affiliated to the ALMAnaCH 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 interests 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.

Notes

1. See https://htr-united.github.io/document-your-data.html (27/04/2023).
2. The letters stand for Findable, Accessible, Interoperable and Reusable. For more information, see https://www.go-fair.org/fair-principles/ (27/04/2023).
3. See https://htr-united.github.io/catalog.html (27/04/2023).
4. See https://github.com/HTR-United (27/04/2023).
5. See in particular https://github.com/HTR-United/htr-united/blob/master/htr-united.yml (27/04/2023).
6. See https://github.com/HTR-United/schema (27/04/2023).
7. See https://htr-united.github.io/actions.html (27/04/2023).
8. See https://github.com/HTR-United/template-htr-united-data-repo (27/04/2023).
9. For more details about these tools, see https://htr-united.github.io/tools.html (27/04/2023).

Bibliography

Chagué, Alix / **Clérice, Thibault** (2022b, June 23). *Sharing HTR datasets with standardized metadata: The HTR-United initiative*. Documents anciens et reconnaissance automatique des écritures manuscrites. https://inria.hal.science/hal-03703989

Kiessling, Benjamin / **Tissot, Robin** / **Stokes, Peter** / **Stökl Ben Ezra, Daniel** (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, *2*, 19–19. https://doi.org/10.1109/ICDARW.2019.10032

Muehlberger, Guenter / **et al.** (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, *75*(5), 954–976. https://doi.org/10.1108/JD-07-2018-0114