HTR and the crowd. A hybrid approach to transcribing civil records from Curaçao

DH Benelux 2024

Thunnis van Oort, Björn Quanjer, Lisa Hoek, Erik Tjong Kim Sang, Coen van Galen

The digital revolution has given historians access to vast collections of scans of civil registrations. But how can we efficiently digitize these registrations to enable analysis of the rich data enclosed in these records? Crowdsourcing and Handwritten Text Recognition (HTR) promise great rewards for large-scale digitization, but each of the methods has its disadvantages. Collaborations with citizen scientists can create highly reliable transcriptions, but this approach is time-intensive. HTR, on the other hand, can process entire texts quickly, but is less reliable because of its larger margin of error, especially if you want to create structured data with specific entities such as names, dates, occupations, et cetera.

In this paper we present our pipeline for a hybrid method where we combine citizen science, HTR and entity recognition to transcribe death certificates from the Caribbean island of Curaçao from the 19th and early 20th century. The main pipeline stages include: page layout analysis and HTR using Transkribus, entity recognition using large language model GPT-4 , human transcription of person names on the certificates, human check of both human and machine input, and post-processing of the data. The pipeline performs at varying accuracy levels while identifying different entity types: around 90% for sex and marital status of the deceased persons, , around 80% for certificate dates and deceased ages, around 50% for professions but only around 30% for person names. The latter performance is the main reason for the pipeline containing an extra human check of the machine analysis.

Our main finding is that the current state of the art of combined automatic HTR and entity recognition is not good enough to satisfy our data quality needs. But we believe that by integrating the automatic methods in our existing citizen scientist-based digitization process, as an additional annotator, we can speed up the process. We are currently testing processing thousands of documents using the pipeline and hope to present the first results at the conference.

References

Lisa Hoek, "Extracting Entities from Handwritten Civil Records using HTR and RegExes". Master's thesis, Radboud University Nijmegen, 2023

Rick Mourits, M. Prats López, Thunnis van Oort, Wessel Ganzevoort and Coen van Galen, "Engaging the Crowd: Citizen Science for Historical Demography". European Social Science History Conference (abstract), Gothenburg, Sweden, 2023.

Erik Tjong Kim Sang, "REE-HDSC: Recognizing Extracted Entities for the Historical Database Suriname Curacao". Technical Report, Netherlands eScience Center, 2023, DOI: 10.48550/arXiv.2401.02972. https://arxiv.org/abs/2401.02972

Thunnis van Oort, Björn Quanjer, Lisa Hoek, Matthias Rosenbaum- Feldbrügge, Coen van Galen, and Jan Kok, "Defeating the Haystack. Combining HTR and entity recognition to reconstruct populations of the past: the case of Curaçao". 5th Conference Of The European Society Of Historical Demography, Nijmegen, The Netherlands, 2023.