

HTR and the crowd

A hybrid approach to transcribing civil records from Curaçao



Thunnis van Oort*, Erik Tjong Kim Sang*, Björn Quanjer, Lisa Hoek, Coen van Galen
DH Benelux 7 June 2024

Radboud University



RADBOD GROUP FOR
HISTORICAL DEMOGRAPHY
AND FAMILY HISTORY

HISTORICAL DATABASE OF SURINAME & THE CARIBBEAN

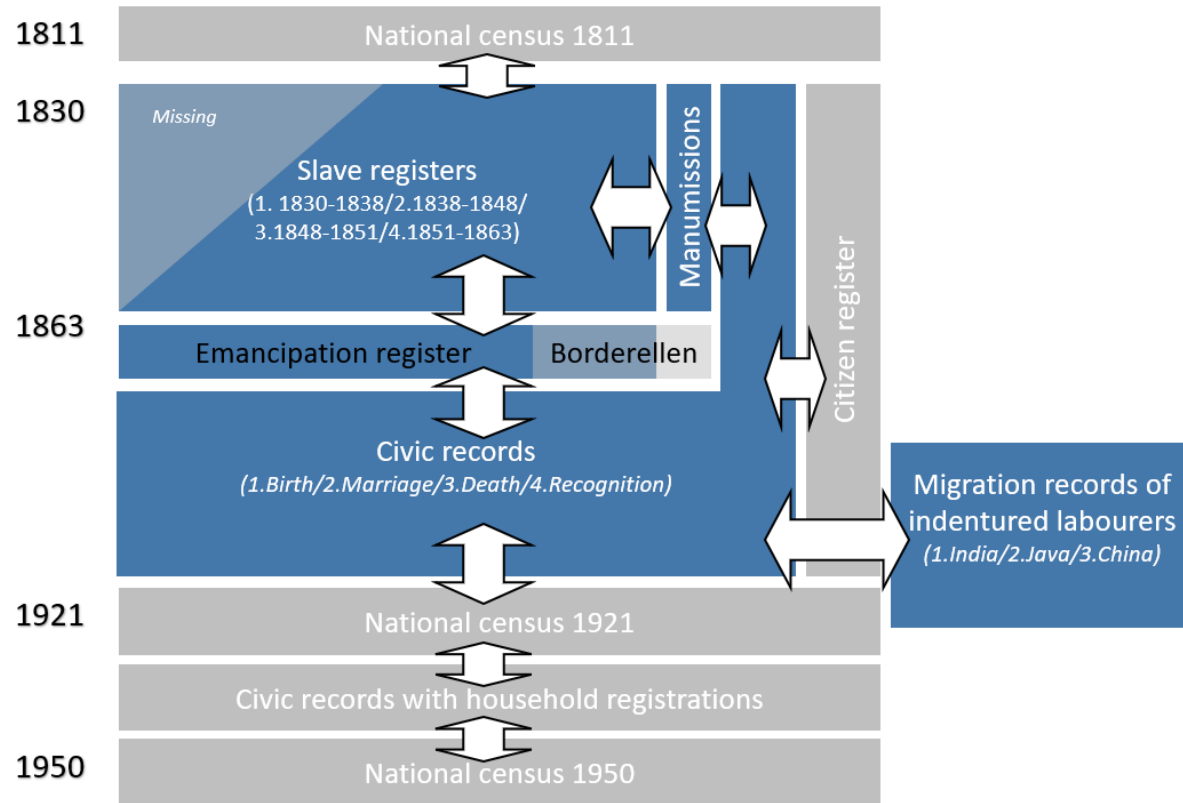


The **aim** of the HDSC is to build a historical database of the entire population of Suriname and former Dutch Antilles (1830-1950) and to make these databases freely available to researchers and the public.

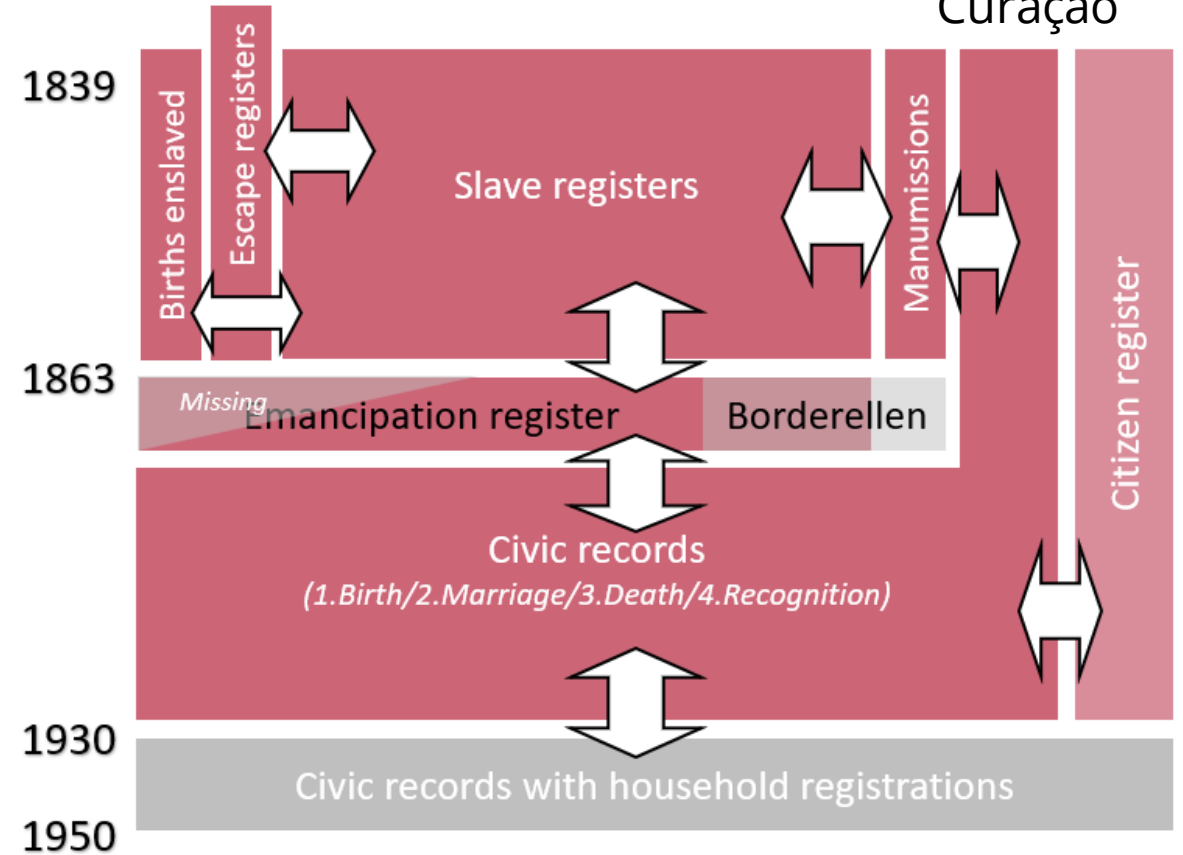
LINKING HISTORICAL POPULATION DATA

HDSC Project

Suriname



Curaçao



Citizen science: data entry civil records

Surinaamse geboortekanten 1869-1900 Handtekening Beheerder Radboud uitloggen

Voer alle informatie uit de geboortekante in in onderstaand formulier.

Invoeren **Instructies**

folio nummer * staat rechtsboven

datum aangifte * dd-mm-jj

plaats aangifte * lettelijk overnemen (ook Paramaribo)

voornaam aangever * voornamen of voornamen

voorvoegsel aangever * eventueel voorvoegsel

achternaam aangever * achternaam (zonder voorvoegsel)

leeftijd aangever * leeftijd in cijfers

beroep aangever * beroep aangever, lettelijk overnemen

woonadres aangever * lettelijk overnemen, voorvoegsel als 'aan den' weglaten

overige informatie aangever * indien expliciet benoemd

handtekening aangever? *

getuige

voornaam getuige * voornaam of voornamen

voorvoegsel getuige * eventueel voorvoegsel

achternaam getuige * achternaam

leeftijd getuige * leeftijd in cijfers

beroep getuige * beroep getuige, lettelijk overnemen

woonadres getuige * lettelijk overnemen, voorvoegsel als 'aan den' weglaten

handtekening getuige? *

getuige

voornaam getuige * voornaam of voornamen

voorvoegsel getuige * eventueel voorvoegsel

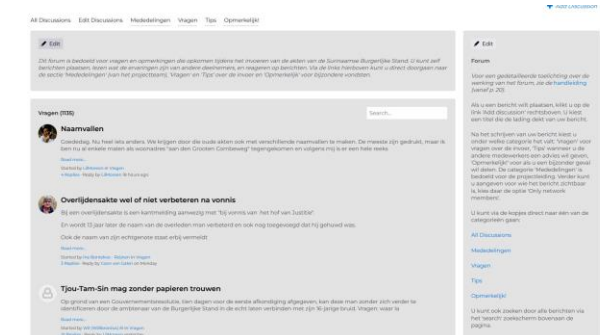
achternaam getuige * achternaam

leeftijd getuige * leeftijd in cijfers

beroep getuige * beroep getuige, lettelijk overnemen

woonadres getuige * lettelijk overnemen, voorvoegsel als 'aan den' weglaten

- Platform: hetvolk.org
- Forum hdsc.ning.com
- Each certificate entered 2x + 1x checked (controllers)
- ~300k certificates... (x 3 tasks)
- Since Sept. 2021: ~540k tasks completed



Citizen Science vs. (Hybrid) HTR

CITIZEN SCIENCE

PROS

- High quality results
- Citizen involvement
- Direct contact with user group

CONS

- Recruitment & coordination takes time & effort (= €)
- Time-intensive
- Production decrease over time

(HYBRID) HTR / ENTITY EXTRACTION

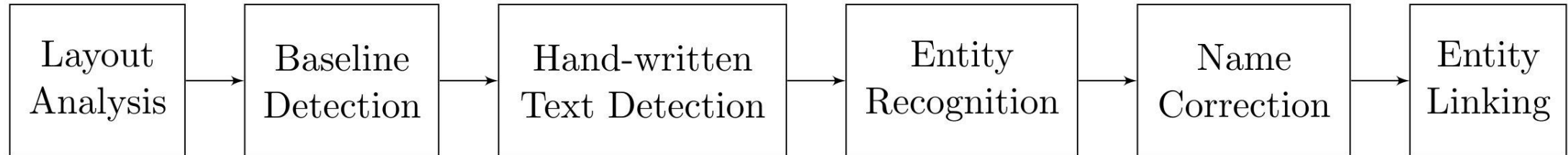
PROS

- Can handle large volumes in brief timespan
- Improvement over time through increased training material
- Transferable to other projects?

CONS

- Lower quality results
- Lack of training data
- Risk of citizen scientists losing motivation
- Costs for processing each document
- During pilot: still time-intensive
- Requires technical expertise

Proposed automated pipeline



1. Layout analysis: determine page layout: how many text columns?
2. Baseline detection: find the lines of text
3. Hand-written Text Detection: get the text of the document
4. Entity Recognition: find the names and dates in the text
5. Name correction: integrate name corrections mentioned in text
6. Entity Linking: link names of identical persons

For step 1-3 we use the program <https://readcoop.eu/transkribus>

N^o 1
Verbeteringen.

Heden den Negen en twintigsten January des Jaars
Een duizend acht honderd Een en Dertig

Compareerden voor mij Ambtenaar van den Burgerlijken stand op dit eiland.

De personen van Isaac Johannes Rammelmaan Elvies
junior en Pieter Pletsz

van competenten ouderdom en alhier woonachtig.

* Dewelke aangifte hebben gedaan dat op dit eiland op den twintiensten
ary des jaars Een duizend acht honderd een en dertig
te half vyf ure des morgens overleden is: Jacob Bonaventura
Rammelmaan Elvies

Oud Twee maanden en twee dagen

Geboren te Curacao

op den dertigsten Juny des jaars Een duizend acht
honderd en dertig

zijnde Blank

laatstelyk gewoond te Curacao

Ongehuwd

En is deze behoorlijk geteekend na voorlezing

J. J. Rammelmaan, Ambtenaar

op Curacao dato ut supra.

In stede van den
Ambtenaar voornemen

P. Pletsz

H. Pletsz
k. v. i. e. n. g.

F^o 1
Eerste Veld
M. de Kanter.
Aanteekeningen.

Dere inschrijving is
geschied ingevolge
Publicatie van Dine.
tuur ad^m in Rade del
5-13 January 1831.

Using Transkribus

Transkribus provides two foundation models for recognizing hand-written historical Dutch: Ijsberg and The Dutchess

Both models perform well on the HTR task but less well on layout detection: they fail to identify text column boundaries

We fine-tuned The Dutchess with 278 death certificates from Curaçao to improve both layout detection and HTR

Evaluation of pipeline

Our HTR model (available as Transkribus model 42578) reports a character error rate (CER) of 3.7% (96.3% correct), which sounds good

However, when we checked different parts of the texts, the performance was mixed: 90% correct for dates but only 33% for names

The character error rate is always lower than the word error rate and this is again lower than the error rate for names

Name accuracy results of HTR model retraining

	1830-1839	1840-1849	1920-1929
Fine-tuning with 5,712 names	68%	69%	41%
Replacing unknown names	68%	66%	43%
Removing unknown names	93% (65%)	88% (67%)	73% (44%)

Postprocessing (replacing unknown words with known words) did not always increase performance for the best retraining configuration

High performances can be obtained for part of the data

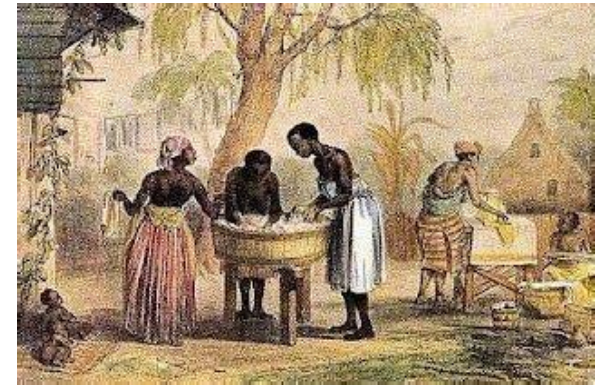
Integration of crowd sourcing and automatic methods

We are currently testing the first combination of crowd sourcing in combination with automatic methods:

- Processing 24,000 death certificates of Curaçao with Transkribus
- Comparing regexes with ChatGPT4 for entity recognition
- Transkribing person names and professions by citizen scientists (sample)
- Integrating machine output in crowd-sourcing pipeline (sample)

We are interested in the time investment of the latest step

Concluding remarks



- Automatic hand-written recognition is a promising method for historical text digitization
- Results are not perfect and require additional manual checks, especially for names
- Effects of using automatic methods on motivation citizen scientists are not yet clear
- Several challenges of the automatic methods remain to be conquered
- Pipeline applicable for other HTR projects? We value input/suggestions for collaboration

Data Publications

- The data are accessible via search interfaces on the national archive websites of Suriname, Curaçao, Aruba and The Netherlands
- The data are published on IISH Dataverse: datasets.iisg.amsterdam/dataverse/HDSC
- See our website: ru.nl/hdsc
- Documentation is available in data papers
e.g. Slavery in Suriname. A Reconstruction of Life Courses, 1830–1863. (2023).
Historical Life Course Studies, 13, 191-211. <https://doi.org/10.51964/hlcs15619>



The screenshot shows the website of the National Archive of Suriname (Nationaal Archief Suriname). The header includes navigation links: Onderzoeken, Studiezaal, Publiek, Actueel, and Over ons. Below the header, there is a breadcrumb trail: Start / Onderzoeken / Personen / Slavenregister. A message states: "Hulp nodig bij het gebruik van het register?" with a "Zoektips" button. The search interface includes a text input field "Vul je zoekterm in..." with a search button "Zoeken" and a dropdown "Uitgebreid zoeken". Below this are filters for (SLAVEN-)NAAM, GESLACHT, MOEDERSNAAM, INSCHRIJFDATUM, and UITSCHRIJFDATUM. The results section shows "Aantal resultaten: 161.790" and "Pagina 1 van 6.472". A table displays the first three results:

(SLAVEN-)NAAM	GESLACHT	NAAM MOEDER	UITSCHRIJFDATUM
Dikie	man	Onbekend	1 juli 1863
Lodewijk	man	Onbekend	1 juli 1863
Casper	man	Onbekend	10 december 1856

The Historical Database of Suriname and Curacao was made possible thanks to the support of our more than 1,000 volunteers and the following partners:



UNIVERSITY
OF CURAÇAO
DR. MOISES DA COSTA GOMEZ



GERDA HENKEL STIFTUNG



PRINS BERNHARD
CULTUURFONDS

For more information on the Historical Database of Suriname and Curacao, see:

ru.nl/hdsc



Radboud University



RADBOUD
GROUP FOR
HISTORICAL
DEMOGRAPHY
AND FAMILY
HISTORY