

Digitizing and Linking Civil Registry Records: Experiences with Transkribus and Death Records of Curaçao

Erik Tjong Kim Sang, Lisa Hoek, Matthias
Rosenbaum-Feldbrügge, Björn Quanjer,
Thunnis van Oort, Coen van Galen

netherlands
eScience center

Monday 22 September 2023

Creating smart solutions for researchers

"What drives me the most is helping others in finding solutions to complex problems."

Hey, I'm Dafne! This is my story



A bit about us

We're an independent foundation with 80+ passionate people working together in the Netherlands' national centre for academic research software.

[Get the whole story](#)

How to read us

Calls for proposals, workshop and job alerts delivered right to your inbox? Sign up for our newsletters and email alerts. We promise not to spam.

[Subscribe now](#)

How to reach us

Science Park 402 (Matrix III)
1098 XH Amsterdam
info@esciencecenter.nl
[+31 \(0\)20 460 4770](tel:+31204604770)

[Ask Google Maps](#)

We're social too! [Twitter](#) [LinkedIn](#) [GitHub](#) [Medium](#) [YouTube](#)

[Privacy policy](#) [Disclaimer](#)

HDSC project, Radboud University Nijmegen

HDSC stands for Historical Database Suriname and Caribbean

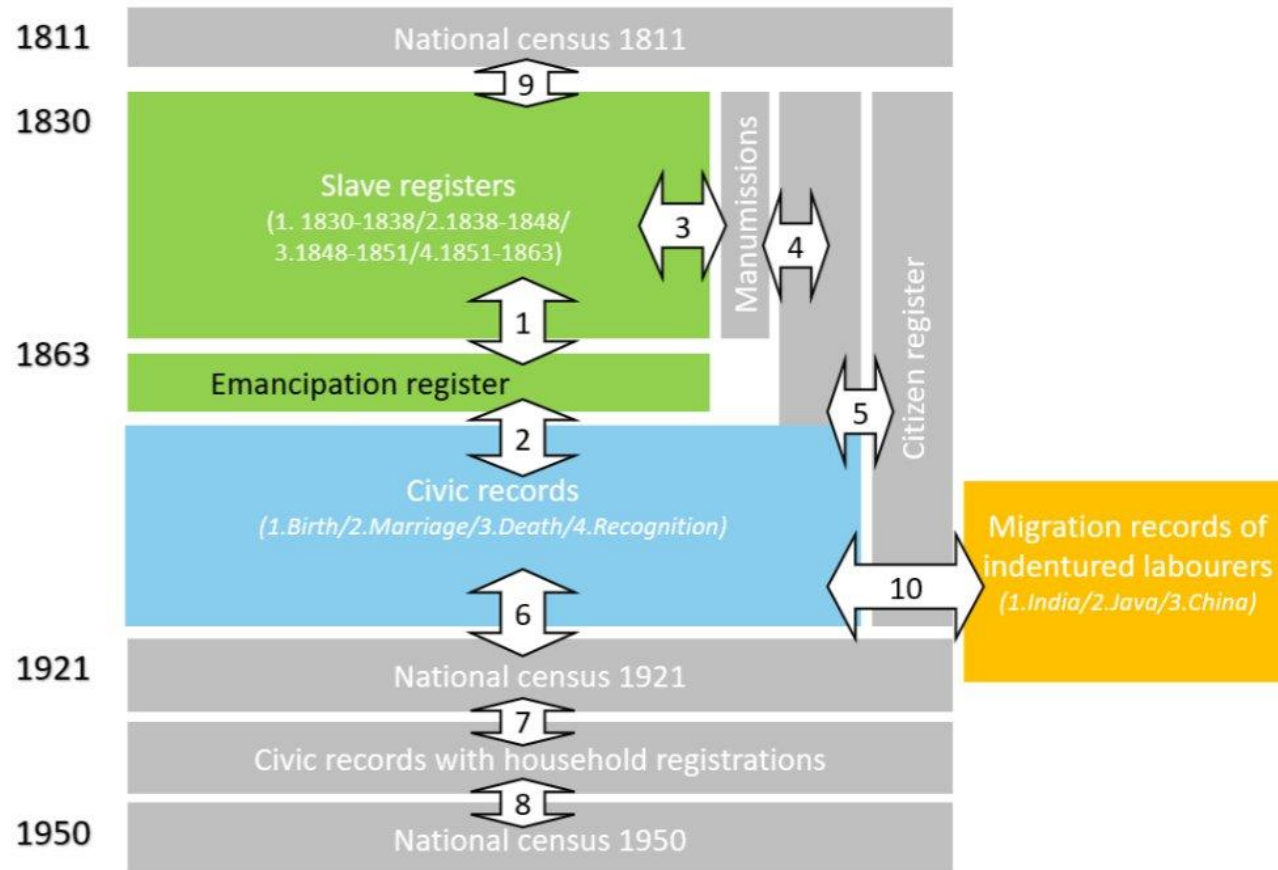
That project aims at collecting and make available data on inhabitants of Suriname and the Caribbean, from the nineteenth century onwards

The populations at time consisted of enslaved people, free people, indigenous people and immigrants, recorded in different registers

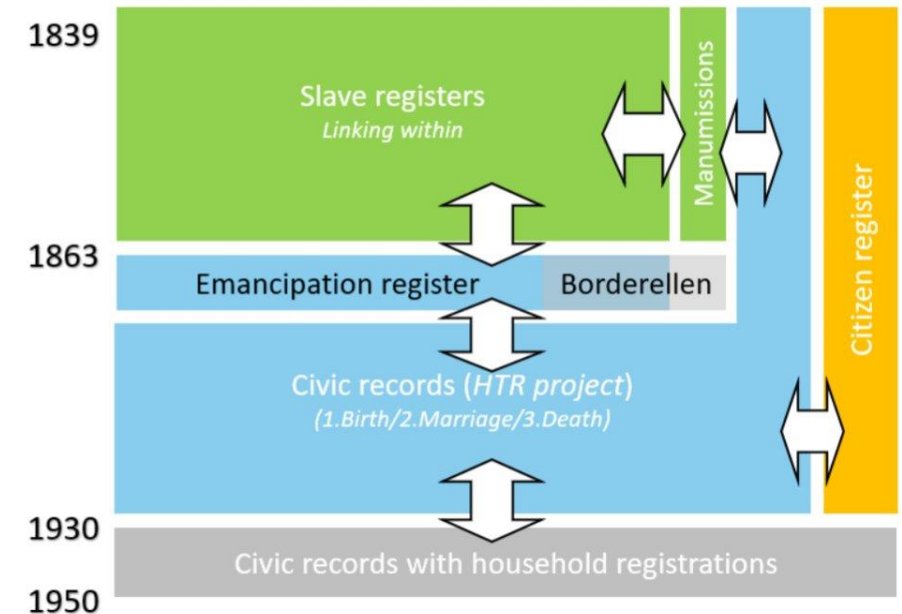
Target groups for data: researchers and genealogists

Overview of civic registers

SURINAME



CURACAO



<https://www.ru.nl/hdsc>

Current practice

Surinaamse Overlijdensakten III (districten 1846-1880)

Handleiding

Op heden den *achtentwintigsten* dag der maand *December*
achttien honderd *zeventig* — is door ons
daniel Cohen & zoon
Ambtenaar van den Burgerlijken Stand in het district *Paramaribo*
Ontvangen een schriftelijke aangifte van
Joannes Fredericus Richardus de Rooy
oud *veertig* jaren, van beroep *planter*
wonende op de plantage *Dordrecht*, mede ondertekend door
Abraham Constantijn Seimans
oud *veertig* — jaren, van beroep *planter*
wonende op de plantage *Dordrecht*
en van *Abraham Josephus Trousfant*
oud *veertig* — jaren, van beroep *Maïchanyje*
wonende op de plantage *Dordrecht*
als getuigen — dat op den *achtentwintigsten*

Controleer de invoer: kies de juiste invoer, corrigeer of vul aan.

Invoeren

Instructions

folio nummer *

staat links- of rechtsboven

datum aangifte *

dd-mm-jjjj

plaats aangifte *

letterlijk overnemen (ook Paramaribo)

schriftelijke aangifte? *

voornaam aangever *

voornaam of voornamen

voorvoegsel aangever

eventueel voorvoegsel

achternaam aangever *

achternaam (zonder voorvoegsel)

leeftijd aangever *

leeftijd in cijfers

beroep aangever *

beroep aangever, letterlijk overnemen

woonadres aangever *

letterlijk overnemen, voorvoegsels als 'aan c

relatie tot overledene *

indien expliciet benoemd: letterlijk overnem

overige informatie
aangever

indien expliciet benoemd

handtekening
aangever? *

getuige

voornaam getuige *

voornaam of voornamen

voorvoegsel getuige

eventueel voorvoegsel

achternaam getuige *

achternaam

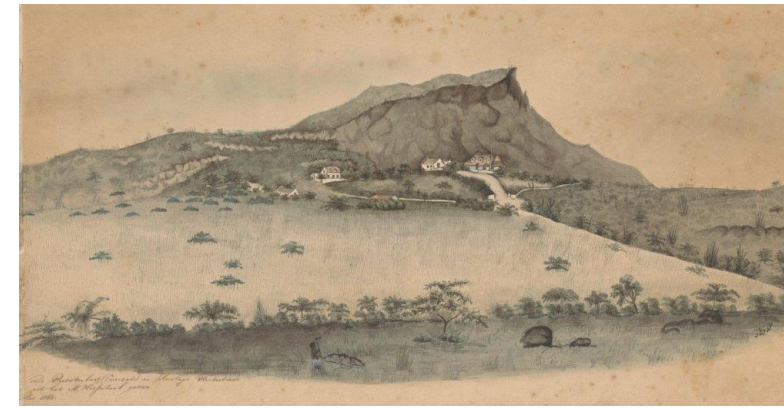
leeftijd getuige *

leeftijd in cijfers

Data from scans of civil registry documents are manually entered in forms by hundreds of volunteers

<https://hetvolk.org>

REE-HDSC sub-project



How can hand-written text recognition (HTR) technology and automatic named entity recognition be integrated in the digitization workflow?

Partners

- Data Science group, Radboud University Nijmegen (Lisa Hoek)
- Netherlands eScience Center (Erik Tjong Kim Sang)

<https://research-software-directory.org/projects/ree-hdsc>

Transkribus

Popular trainable tool for optical character recognition (OCR) and hand-written text recognition (HTR): from images to text

Processing is performed in three steps:

1. Layout analysis of page
2. Baseline detection: find position of lines of text
3. Optical character recognition and hand-written text recognition

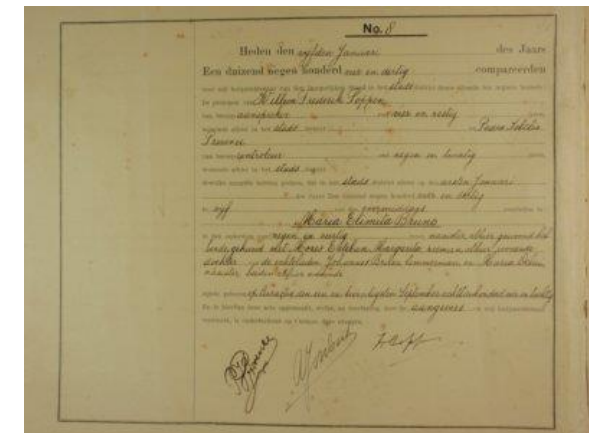
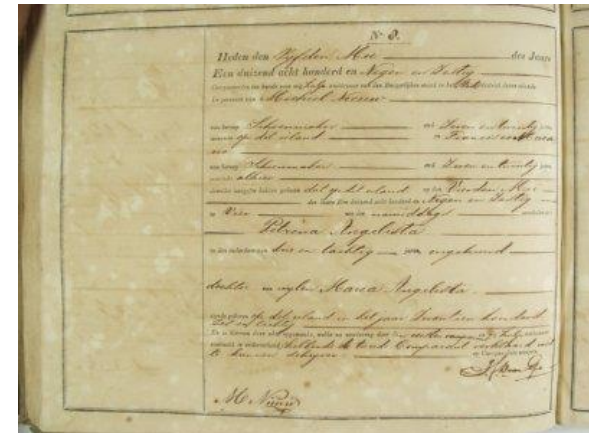


Layout analysis

We found three text formats among the death certificates of Curaçao:

1. Three-column layout (1831-1869)
2. Early two-column format (1869-1933)
3. Late two-column format (1934-1950)

Two layout analysis modules were trained for this corpus (three-column and two-column)



Transkribus evaluation

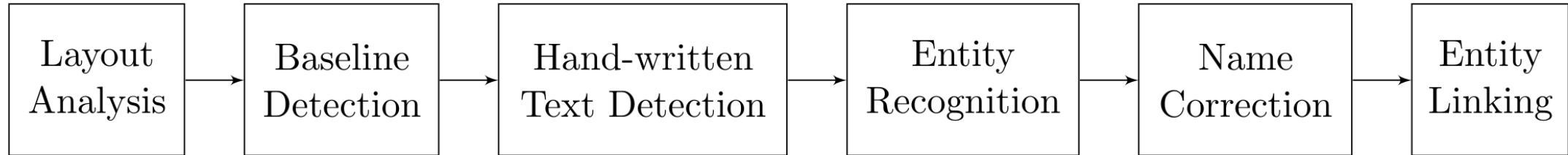
After training on our data, Transkribus achieves a character error rate of about 3% (Hoek, 2023)

However, a disproportional number of errors occurs in the most important parts of the text: the person names

These errors are difficult to recover with post-processing



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: make original deceased name complete
6. Entity Linking: link names of identical persons

Entity recognition

We considered four approaches for entity recognition:

1. Regular expressions only,
2. Arbitrary entity recognition (with a Dutch Huggingface model) plus rules, implemented for deceased name and death date,
3. Specific entity recognition (for example deceased name): requires model fine-tuning, not realized yet
4. ChatGPT: give ChatGPT a text and ask it to extract information from it

Evaluation of entity recognition

	Deceased Names			Death Dates		
Method	Found	Exact	Partial	Found	Exact	Corrected
REG only	83%	14%	-	74%	73%	-
ML + REG	95%	17%	55%	76%	37%	58%
GPT3.5	100%	33%	83%	100%	82%	90%
GPT4	100%	34%	81%	100%	84%	91%

Partially correct names: with a Levenshtein distance of up to 3
Corrected dates: taking into account the year in file name

Entity linking

All names of a person appearing in different documents need to be linked to each other

We have linked people with the same name and birth date

Challenges in our data:

1. People regularly change name
2. Birth dates are not always available (registrants/witnesses)
3. Care should be taken when dealing with data from twins

Future work

Improve the quality of the HTR

- Create more training material for Transkribus
- Expand the language model with names, but how?

Entity recognition

- Can the quality of machine learning output be improved?
- Do we want to rely on a tool like ChatGPT?

Entity linking

- When should we make a link and when not?



THE END

netherlands
eSciencecenter