

Recognizing Extracted Entities for the Historical Database Suriname Curaçao

Erik Tjong Kim Sang
Thursday 30 November 2023

netherlands
eScience center

REE-HDSC project

Radboud University started the HDSC project in 2017

That project aims at collecting and make available historical data on inhabitants of Suriname and the Caribbean

The task of the eScience Center is to examine how the required text-to-data conversion can be automated

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



HDSC People

- Matthias Rosenbaum-Feldbrügge, Radboud University Nijmegen, LA
- Björn Quanjer, Radboud University Nijmegen, Researcher
- Thunnis van Oort, Radboud University Nijmegen, Researcher
- Lisa Hoek, Radboud University Nijmegen, Data scientist
- Coen van Galen, Radboud University Nijmegen, Project Manager
- Erik Tjong Kim Sang, eScience Center, Research Software Engineer

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

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 Rammelman Elsiever
Junior en Pieter Plets

van competenten ouderdom en alhier woonachtig.

De welke aangifte hebben gedaan dat op het eiland op den tweeden Janu-
ary des Jaars Een duizend acht honderd een en dertig
te half vyf ure des morgens overleden is: Jacob Bonaventura
Rammelman Elsiever

Oud Les maanden en twee dagen

Geboren te Curacao

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

zijnde Blank

laatstelijk gewoond te Curacao

Ongehuwd

En is deze behoorlijk geteekend na voorlezing

J. J. Rammelman, Ambtenaar

op Curacao dato ut supra.

In stede van den
Ambtenaar voorlezend

P. Plets

P. Plets
H. P. van

F^o 1
Eerste blad.
M. de Kanton.
Aanteekeningen.

Dere inschrijving is
geschied ingewolge
Publicatie van Diss.
tus adm en Rade tot
5-13 January 1831.

Current practice

Surinaamse Overlijdensakten III (districten 1846-1880) Handleiding

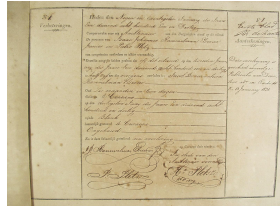
Controleer de invoer: kies de juiste invoer, corrigeer of vul aan.

Invoeren Instructies

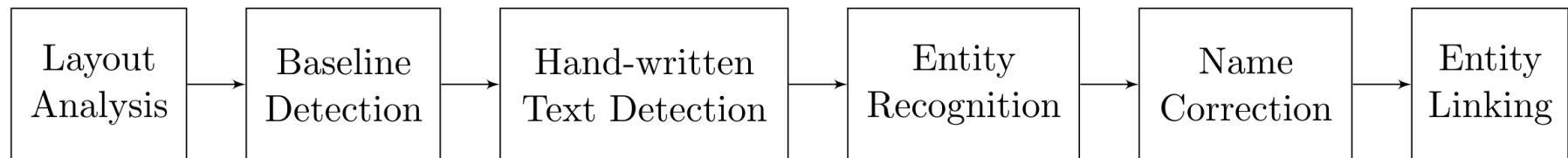
folio nummer *	staat links- of rechtsboven
datum aangifte *	dd-mm-ijij
plaats aangifte *	letterlijk overnemen (ook Paramaribo)
schriftelijke aangifte? *	
voornaam aangever *	voornaam of voornamen
voorvoegsel aangever	eventueel voorvoegsel
achternaam aangever *	achternaam (zonder voorvoegsel)
leeftijd aangever *	leeftijd in cijfers
beroep aangever *	beroep aangever, letterlijk overnemen
woonadres aangever *	letterlijk overnemen, voorvoegsels als 'aan t
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>



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: identify 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>

Evaluation of pipeline

The hand-written text module reports a character error rate (CER) of 6.0% (94% correct), which sounds good

However, when we tried to identify names, 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



Improving hand-written text recognition (HTR)

We have three ideas for improving the quality of person names by hand-written text recognition:

1. Additional training with only hand-written person names
2. Post-processing of the output of the HTR module
3. Automatic quality assessment of HTR output

For training we have tens of thousands of already identified names

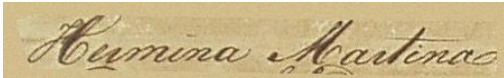
Name accuracy results of HTR model retraining

	1830-1839	1840-1849	1920-1929
Before post-processing	68%	69%	41%
After post-processing	68%	66%	43%
Quality estimation	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

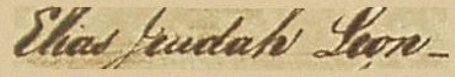
HTR examples



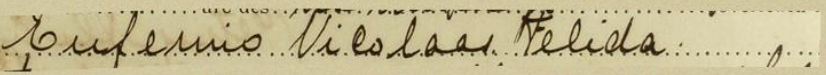
Hermina Martina



Henrietta Wilhelmina Fillis



Maria Elias Junda Son



Lucina Hilda

Name found	Evaluation
Hermina Martina	Correct
Henrietta Wilhelmina Fillis	Near miss
Maria Elias Junda Son	Hallucination
Lucina Hilda	Complete miss



Concluding remarks

- Hand-written text recognition can be used for processing old documents
- This technique does not perform perfectly, in particular not for names
- Fortunately new recognition models can be trained for our data
- We trained a new model for Dutch civil records: 68% name accuracy
- Post-processing did not improve the scores
- Accuracies close to 90% could be reached for part of the documents (66%)
- Some sets of documents proved to be harder for processing than others



THE END

netherlands
eScience center