

# Golden Agents: Detectie van entiteiten in boedelinventarissen

Maarten van Gompel, Bram Buitendijk, Leon van Wissen, Harm  
Nijboer, Menzo Windhouwer

23 juni 2022

# Introductie

- ▶ Boedelinventarissen beschrijven de huisraad, activa en passiva.
  - ▶ geen onroerend goed
  - ▶ nav huwelijk, faillissement, nalatenschap etc. . .
- ▶ Deze willen we makkelijker ontsluiten voor onderzoekers
  - ▶ lange onderzoekstraditie (archeologisch, sociaal-historisch, etc..)

**Onze doelstelling:** Hoe kunnen we *automatisch* namen van personen, locaties, en vooral objecten herkennen in boedelinventarissen?

# Dataverwerkingspipeline (1)

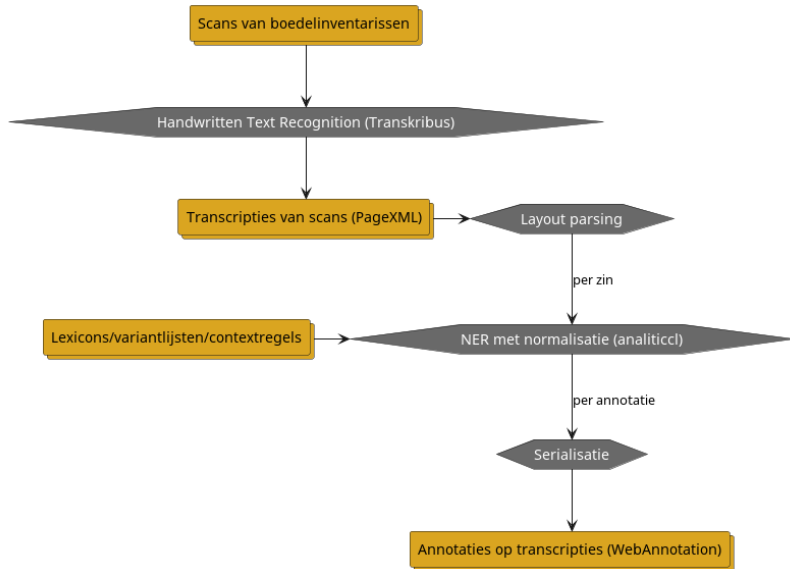
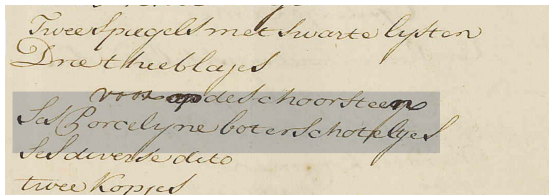


Figure 1: Pipeline

# Uitdagingen

- ▶ De HTR output bevat allerlei spellingsvariatie:
  - ▶ door HTR-fouten
  - ▶ door diachronische spellingsvariatie en gebrek aan standaardisatie indertijd door spatiëeringsfouten (splits/run-ons)
- ▶ Normale entiteitherkenning werkt hierdoor niet goed
  - ▶ Standaard NER-modellen zijn niet getraind op objecten



Tien Printe bortjes  
Twee spiegels met swarte Ejsten  
Drie theeblejes  
voor op deschoorsteer  
**Tes Porcelijne boterschotpjes**  
ses diverse dite  
twee kopjes

Figure 2: HTR Voorbeeld

# Onze strategie

## Wat hebben we?

- ▶ Een flink aantal lexicons/thesauri met namen van personen, locaties en objecten (diverse bronnen)
- ▶ Met name de objecten hebben we handmatig verrijkt:
  - ▶ bron
- ▶ INT Historisch Lexicon; gecureerde lijst die historische varianten koppelt, tevens geschikt als *achtergrondlexicon*
  - ▶ helaas niet openbaar beschikbaar ivm beperkte rechten

## Aanpak

- ▶ We doorzoeken boedelinventarissen op termen uit deze lexicons, rekening houdend met:
  - ▶ spellingsvariatie; zoek de begrippen in de lexicons die het meeste op de aangetroffen vorm lijken
- ▶ combinatie van named entity recognition en tekstnormalisatie in één

## Dataverwerkingspipeline (2)

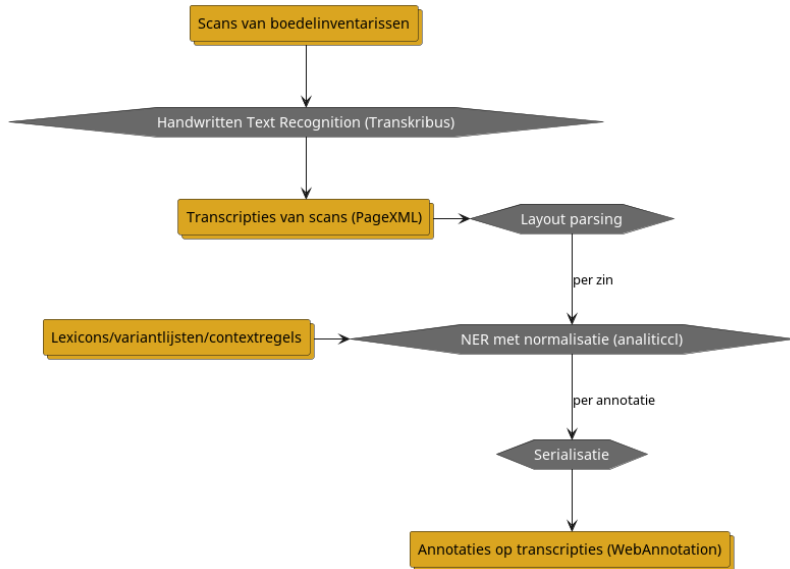


Figure 3: Pipeline

# Analiticcl

- ▶ Software voor spellingscorrectie en tekstnormalisatie
  - ▶ koppelt woorden en frasen aan varianten
  - ▶ leest en genereert *variantenlijsten*
  - ▶ doet zowel *correctie* als *detectie*
  - ▶ verschillende afstandsmetrieken (waaronder Damerau-Levenshtein)
  - ▶ *schaalbaarheid*: gaat efficiënt om met grote zoekruimten en zoekafstanden
  - ▶ lexicons kunnen *frequentieinformatie* bevatten
  - ▶ *context*: weegt contextinformatie mee d.m.v. taalmodellen of opgestelde contextregels
  - ▶ deze contextregels maken een soort tagging mogelijk
  - ▶ *command-line tool* en *library*: geïmplementeerd in Rust, met Python bindings
- ▶ Herimplementeert en bouwt voort op kernideeën van Martin Reynaert (TICCL)
- ▶ Technisch-inhoudelijke presentatie:  
<https://diode.zone/w/kkrqA4MocGwxyC3s68Zsq7>

# Voorbeeld Analiticcl

```
proycon@mhyas ~W/analiticcl <master> — 14:25:18 - Tue 18 May — #1
$ target/release/analiticcl query -1 --alphabet ~W/analiticcl/examples/simple.alphabet.tsv --lexicon ~W/analiticcl/example
s/nld.aspell.sonarfreq.lexicon
Initializing model...
Loading lexicons...
Building model...
Computing anagram values for all items in the lexicon...
- Found 222908 instances
Adding all instances to the index...
- Found 209099 anagrams
Creating sorted secondary index...
Sorting secondary index...
- Found 60 anagrams of length 2
- Found 545 anagrams of length 3
- Found 2 anagrams of length 33
- Found 1 anagrams of length 34
- Found 1 anagrams of length 35
- Found 1 anagrams of length 38
Adding ngrams for simple language modelling...
Querying the model...
(accepting standard input; enter input to match, one per line)
huys
huys    huis    0.6395655036208032    huls    0.6066491112574062    huns    0.598968619705947    h
ups     0.5978714066271671    huts    0.5978714066271671    hu      0.5402677199912223    huur0
.5243581303489138    buis     0.47580645161290325    huil    0.4409699363616415    huid    0.436
5810840465219
huysraat
huysraat    huisraad    0.6505376344086021    humoraal    0.4383640552995392    huisbaas    0
.4182027649769585
huwelyck
huwelyck    huwelijk    0.8319892473118281    ouwelijk    0.43847072879330945
vergt n
vergt n    vergen    0.860215053763441    vergun    0.7004044589128934    vergroten    0.6857715958041498    v
ergt     0.6243957778435435    vergeten    0.6204991614876196    vergaten    0.619019433757522    v
ergoten    0.619019433757522    verven     0.6104370129229555    verdun     0.6000789188122719    veree
n        0.586761369241393
4> mhyas> 1 target/release/analiticcl:analiticcl 0 0.67, 0.82, 0.60 14:26:22
```

Figure 4: Voorbeeld, blauwe regels zijn input



# Lexiconcuratie

- ▶ Kwaliteit is erg afhankelijk van de kwaliteit van de input (lexicons, variantlijsten)
- ▶ Handmatige **lexiconcuratie** om tot een lijst 'boedeltermen' te komen. Focus op objecten, maar ook categorieën die helpen deze vindbaar te maken (denk aan materialen, telwoorden en andere eigenschappen).
- ▶ Belangrijke rol voor het achtergrondlexicon (INT Historisch Lexicon)

# Referentiedata en annotatie

Om te kunnen evalueren hebben we referentiedata nodig.

- ▶ Een klein aantal boedelinventarissen zijn handmatig geannoteerd, zowel op categorie (persoon, locatie, object) als op tekstnormalisatie
- ▶ Een annotatieomgeving (gebaseerd op Recogito-JS) is speciaal hiervoor ontwikkeld:
  - ▶ Source:  
<https://github.com/knaw-huc/golden-agents-htr/tree/master>
- ▶ Annotatoren hebben een eerste output van analiticcl gecorrigeerd en aangevuld om tot een ground-truth te komen
  - ▶ dank ook aan Jirsi Reinders & Judith Brouwer
- ▶ Op deze development-set hebben we verdere parameters getest en input lexicons verbeterd
- ▶ Een evaluatietool vergelijkt systeemoutput (analiticcl) met de referentiedata en berekent Precisie, Recall en F1.

# Voorbeeld Annotatietool

## Golden Agents: Annotation Evaluation (v2022.03.23) (?)

Text: ☐ ☒ A25555000113

Checked: Harm ☐ Jirsi ☒

Judith ☐

[Save annotations](#)

Tag Legend: [\(ambiguous\)](#) | [firstname](#) | [familyname](#) | [person](#) | [occupation](#) | [material](#) | [property](#) | [object](#) | [picture](#) | [animal](#) | [category](#) | [quantifier](#)

[Page in Stadsarchief](#)

t  
18  
24:  
f  
Inventaris en taxatie van de [gereetslappen](#)  
door den versuerder overgelevert en door de  
guerdens vande voorsz. brouwerij gereedschap  
en bekennen bijde ondertekener  
haer ontfangen zijn, onder c  
huer sedul vermeldt Staet, Na  
de [nieuwe waterschuijt](#) cost  
[een entachtigh gulde sacht](#)  
-  
Segge aengenomen voor  
-  
f 1481:8  
de [oude waterschuijt](#) aengenomen  
voor [hondert en vijftigh guld](#) Segge f 150:

gereedschap

gereedschappen

Add a reply...

object

Add tag...

[Cancel](#)

[Ok](#)

# Voorbeeld resultaten

Muylen

f2 28

in een klijn Vast kasje

Een Doosje meediverse sakjes Knoope

Eenige Rommeling

Twee wolle dam dte borstrokken

Twee wolle en eenig Catoene deeken

Een kessen en een klijnder dito

Drie Matrassen

Eenonde Velthafel

Een Vengel kooijje

Eenkleere bakje

Een Tuweel Koffertje nevens eenige Rommeling

op 't Comptoire

Een Bed Peulerenses Kussens

Eenhhoene deken

Twee Materessen

Een klijn wit deekenje

Een Vrouwe Jak en eendito borstrek

Twee Gryne Rokken

In het voorhuijs

Eerstoombank ende winkel plakke

Eenspeecel met eenswartelijst

Een gescheldert gemak Koffertje

# Evaluatie

## Classificatie & Normalisatie

Metriek	#out	#ref	Precisie	Recall	F1
Objecten	448	584	0.498	0.382	0.432
Personen	219	130	0.612	0.638	0.625
Locaties	14	26	0.143	0.077	0.100
Vertrekken	33	33	0.455	0.455	0.455
<b>Totaal</b>	-	-	0.525	0.435	0.475

## Alleen Classificatie

Metriek	Precisie	Recall	F1
Objecten	0.681	0.522	0.432
Personen	0.685	0.714	0.699
Locaties	0.214	0.115	0.150
Vertrekken	0.455	0.455	0.455
<b>Totaal</b>	0.662	0.550	0.601

## Referenties

- ▶ **Analiticcl:** <https://github.com/proycon/analiticcl>
- ▶ **Golden Agents HTR Repo:**  
<https://github.com/knaw-huc/golden-agents-htr>
  - ▶ Bevat: annotatietool, evaluatietool, NER wrapper tool, gecureerde lexicons, alle experimenten (uitvoer, evaluatie logs, sparql queries)

### Publicaties:

- ▶ Reynaert, Martin. (2004) Text induced spelling correction. In: Proceedings COLING 2004, Geneva (2004).  
<https://doi.org/10.3115/1220355.1220475>
- ▶ Reynaert, Martin. (2011) Character confusion versus focus word-based correction of spelling and OCR variants in corpora. IJDAR 14, 173–187 (2011).  
<https://doi.org/10.1007/s10032-010-0133-5>