



Toelichting op het doel van deze SBIR-Innovatiecompetitie

Welke archiefstukken worden getranscribeerd?

Voor het Nationaal Archief gebruiken we scans van de 17^e en 18^e eeuw van de VOC. Dat zijn de 'Overgekomen brieven en papieren' gestuurd aan de Kamer Zeeland en archiefstukken van de Raad van Justitie in Batavia. (Toegangsnummer 1.04.02, Inventarisnummers¹ 7527-9540. Aantal scans: 1.046.951)

Van het Noord-Hollands Archief gebruiken we 19^e eeuwse archieven van de notarissen te Haarlem. (Toegangsnummers 1617, inventarisnummers 1593-1805; en 1972, 1-813. Aantal scans: 420.000)

Het aantal scans wordt aangevuld tot 2 miljoen met notariële archieven van overige Regionaal Historische Centra in de provincie.

Welke datasets zijn beschikbaar?

We hebben een export gemaakt uit het transcriptie platform Transkribus van drie inventarisnummers van de VOC en drie inventarisnummers van het notarieel archief. De export is als dataset beschikbaar en heeft de volgende naam:

- scan-page.export_job_599574

Let op! De export bevat slechts een fractie van de scans die in het kader van dit project worden ontsloten. Zo ontsluiten we van het archief van de VOC 2013 inventarisnummers. Wat weer een fractie is van het totale archief van de VOC bij het Nationaal Archief, dat in totaal 14927 inventarisnummers is en bijna 1,5 km aan plankruimte inneemt.

De export bevat voor elk inventarisnummer een mapje. In dat mapje zitten de scans (JPG) en de transcriptie data (PAGE XML) van dat inventarisnummer.

Dit zijn ook de bestanden waar de functionaliteit voor zoeken en tonen mee moet werken, met het verschil dat de functionaliteit voor zoeken en tonen de afbeeldingen via de International Image Interoperability Framework (IIIF) server ophaalt.

De bestandsnaam van de twee bestanden zijn het zelfde, met uitzondering van de extensie. Zie bijvoorbeeld: NL-HlMnHA_1617_1800_0004.jpg
NL-HlMnHA_1617_1800_0004.xml

Hier staat: Nederland-Noord Hollands Archief_toegangsnummer 1617_inventarisnummer
1800_volgnummer 0004.bestandsformaat

Om voor deze opdracht makkelijker bij de tekst van de transcripties te kunnen, is de volgende dataset beschikbaar:

- txt.export_job_608482.

Het bevat per inventarisnummer een TXT bestand met alle tekst op de scans van dat inventarisnummer.

¹ Een inventaris is een archiefstuk, bijvoorbeeld een gebonden deel of losbladig pak papier.

Wat zijn de voornaamste uitdagingen?

Oerwoud aan transcriptiedata

Als een onderzoeker in de gigantische hoeveelheid transcriptiedata zoekt, hoe kan hij zo goed mogelijk worden geholpen? Kunnen er slimme tools en A.I. worden ingezet om de transcriptiedata te duiden?

Als de onderzoeker bijvoorbeeld zoekt op het trefwoord 'Batavia' dan heeft hij niet alleen te maken met transcriptiefouten (zie ook de alinea 'Transcriptiefouten') maar ook met veel zoekresultaten, wel 222 in een beperkte dataset van drie inventarisnummers van de VOC, zie tabel 1. Het probleem wordt nog omvangrijker als binnen dit project alle 2013 inventarisnummers worden ontsloten. Dit wordt niet meer te overzien als we in de toekomst het hele VOC archief met 14927 inventarisnummers gaan ontsluiten.

Wellicht wil de onderzoeker wel heel specifiek iets weten over het schip met de naam Batavia of het kasteel in Batavia. Hoe kan hij dat dan vinden? Uit onderstaande tabel 1 met een aantal zoekvragen en de zoekresultaten is duidelijk dat 'kasteel in Batavia' niet tot het gewenste zoekresultaat leidt.

zoek	resultaat
batavia	222
datavia	20
patavia	10
bataria	4
kasteel in batavia	0
casteel in batavia	0
Int casteel batavia	2
batavia in't casteel	1
batavia int casteel	1
batavia in 't casteel	4
casteel tot batavia	2
batavia ant casteel	1
iut casteel batavia	1
des casteels batavia	1
IntCasteel batavia	1
.Batavia opthijlant groot Java IntCasteel	1

Tabel 1

Hoe kan de onderzoeker beter worden geholpen in zijn zoektocht door extra context te bieden? Kan A.I. de onderzoeker helpen een weg te banen door de transcriptiedata?

Koppeling met de toegangen

In een toegang wordt een archief beschreven. Als de informatie in de toegang wordt gekoppeld aan de transcripties kan dit worden gebruikt om de zoekresultaten te filteren. De toegangen worden beheerd in het collectiebeheerssysteem Mais Flexis. Een export van de belangrijkste toegangen van de archieven in dit project zijn als dataset beschikbaar:

Nationaal Archief

- NL-HaNA_1.04.02.ead.pdf

Noord-Hollands Archief

- 1617.pdf
- 1972.pdf

Op basis van de bestandnaam van de transcriptie kan het nummer van de toegang en het nummer van het inventarisnummer worden bepaald. Zie voor een uitleg hierover de alinea 'Welke dataset is beschikbaar?'.

Bijvoorbeeld, in dit project hebben we een klein deel van het archief van de VOC geselecteerd: inventarisnummer 7527-9540. Op basis van de hiërarchische beschrijving van het archief in de toegang kunnen we veel informatie achterhalen van de archiefstukken. Zie hieronder de informatie die onder andere beschikbaar is van inventarisnummer 7528, een van de inventarisnummers die beschikbaar is gesteld in de dataset.

> Deel II, Kamer Zeeland

> Deel II/E, INGEKOMEN STUKKEN VAN GOUVERNEUR-GENERAAL EN RADEN BIJ DE HEREN XVII EN DE KAMER ZEELAND

> Deel II/E.1, Missiven

> 7528-7595, Missiven van gouverneur-generaal en raden en (bij uitzondering) van andere colleges in Batavia aan de Heren XVII of de kamer Zeeland. Met bijlagen. 1684 - 1792 68 banden

> 7528, 1684 mei 31 (generale missive); 1684 mei 31 (particuliere missive) 1684

Zo kunnen we met behulp van de toegang vaststellen dat de 1170 scans van dit inventarisnummer geschreven zijn in 'mei 1684' door 'gouverneur-generaal en raden en (bij uitzondering) van andere colleges in Batavia' en dat de 'particulieren en generale missiven' (ambtelijke brieven) gestuurd zijn aan de 'Heren XVII of de kamer Zeeland'. Hiermee kunnen we antwoord geven op de vragen: wie, wat, waar en wanneer.

Hoe ga je deze informatie gebruiken om de onderzoeker te faciliteren?

Transcriptiefouten

We zijn bezig met het ontwikkelen van een trainingsmodel voor het automatisch transcriberen via machine learning. We zijn nog maar net begonnen om de machine te voeden met correcte transcripties. De huidige trainingsmodellen zijn nog niet zo goed. De transcripties die de machine maakt bevatten daardoor nog veel fouten. We verwachten dat dit zal verbeteren in de aankomende maanden maar er zal ook dan rekening gehouden moeten worden met transcriptiefouten.

Zie afbeelding 1 t/m 4 voor voorbeelden van transcriptiefouten. 'Steenbergen' kan getranscribeerd zijn als 'Heenberg' (afb 1, regel 1) en 'Heen bergen'. 'Contract' kan getranscribeerd zijn als 'contraa' (afb 3, regel 2). Of 'Beverwijk' kan getranscribeerd zijn als 'Biverwijk' (afb 3, regel 28).

In de transcriptie van de circa 600 scans van inventarisnummer 8068 van de VOC komt naast Ternate (264 x), een belangrijke VOC post, ook Sernate (77x) en Tennate (4x) voor. Dit zijn transcriptiefouten.

Het moet daarom mogelijk zijn dat als je 'Steenbergen' zoekt je ook 'Heenberg' en 'Heen bergen' als mogelijk goed antwoord vindt. Hoe ga je dat oplossen?

Spellingsvarianten

Tot in de 19^e eeuw was er geen sprake van een vastgestelde spelling en kon een woord in meerdere spellingsvarianten voorkomen. De spelling was vroeger vaak fonetisch.

Zo kan 'eiland' geschreven zijn als 'eijlant' of 'eijland'. 'Koninginnen' als 'Coninginne' (afb 2), 'Sultan' als 'Sulthan' (afb 2) en 'stuivers' als 'stuijvers' (afb 4, regel 11, 30 en 32). Zie ook: https://nl.wikipedia.org/wiki/Alternatieve_spelling_in_het_Nederlands.

Hoe los je dit op? Kan bijvoorbeeld soundex hier een oplossing?

<https://nl.wikipedia.org/wiki/Soundex>

Oude spelling, oud Nederlands en synoniemen

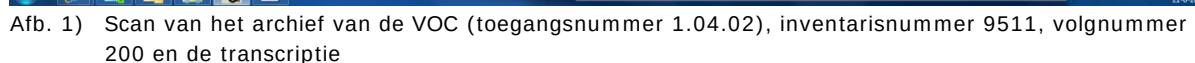
Naast dat we tegenwoordig een andere spelling gebruiken worden er ook andere woorden gebruikt. Zo kennen we tegenwoordig niet meer het beroep 'sjouwerman' (afb 3, regel 20). Is het mogelijk

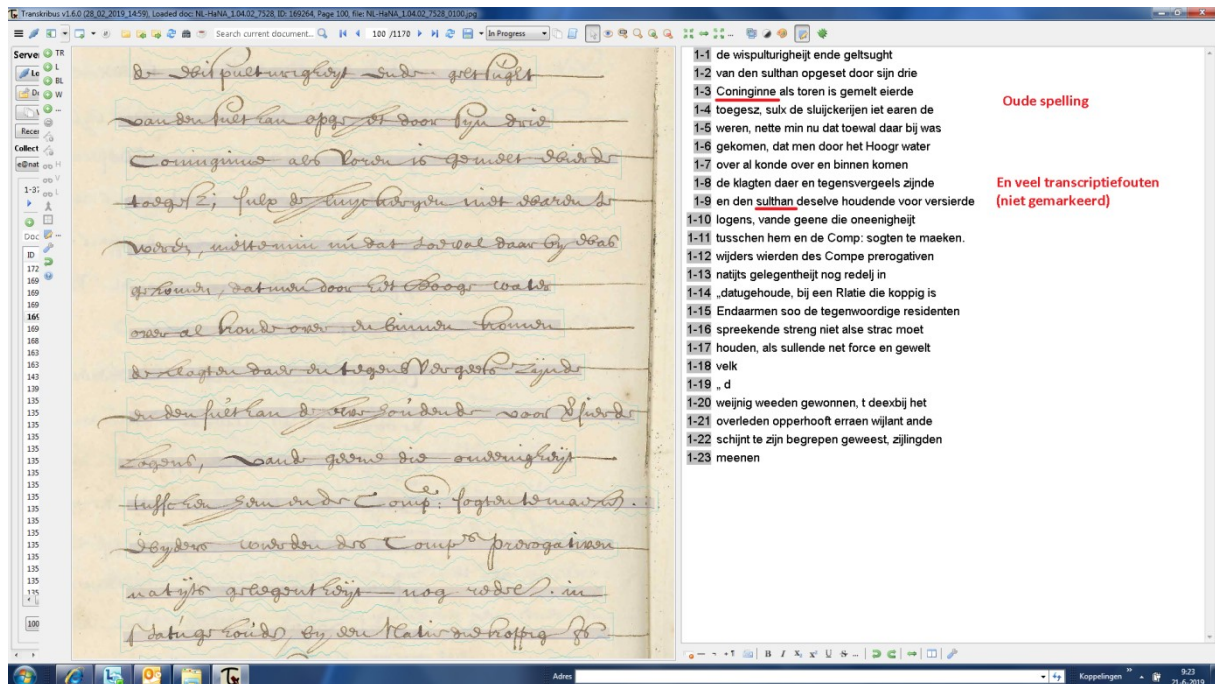
Woordenboek der Nederlandse Taal: <http://gtb.inl.nl/search/?owner=WNT>
 VOC Glossarium: <http://resources.huygens.knaw.nl/vocglossarium>
 Synonym Search: <http://217.104.57.71/cgi-bin/synsearch.cgi?query=schilderij>

Het kan voorkomen dat de lay-out herkenning niet helemaal goed is gegaan. Zo kunnen twee kolommen met tekst niet goed herkend zijn zodat de tekst van deze kolommen door elkaar loopt (afb 3 en 4).

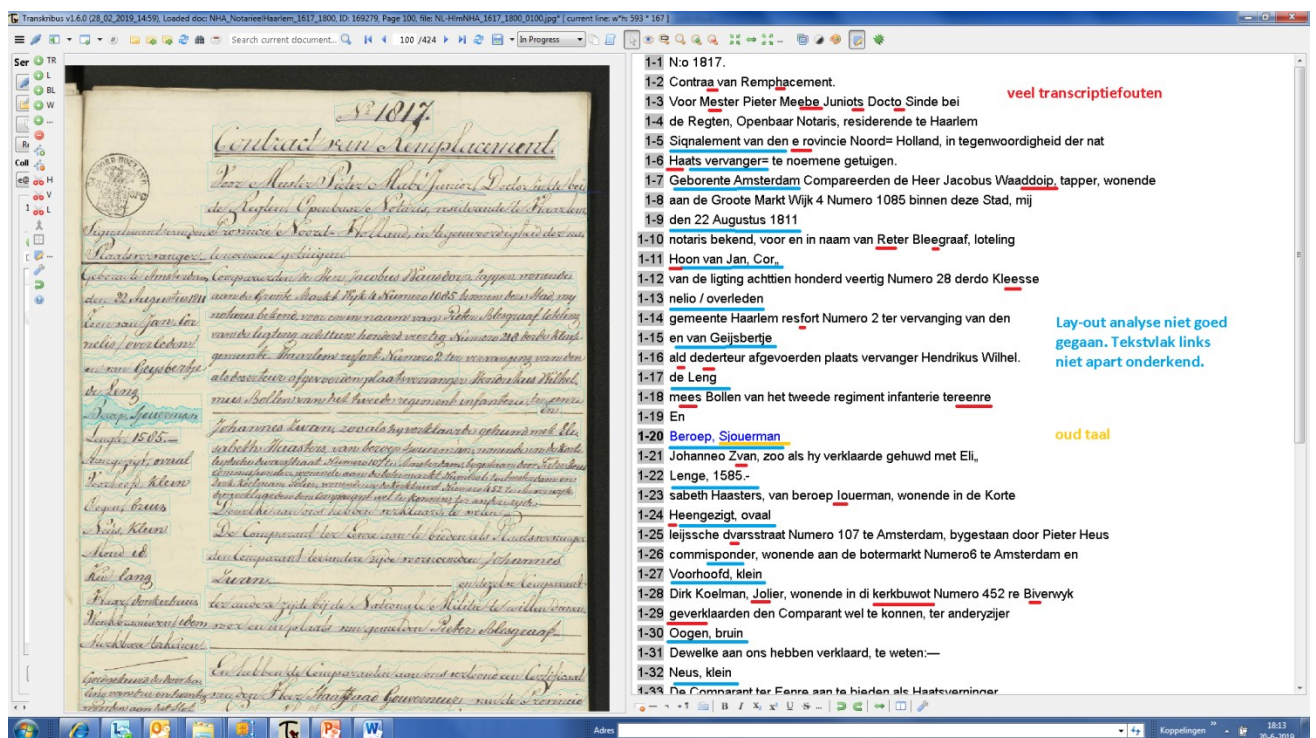
Wellicht heb je voor deze issues een slimme oplossing?

Zie hieronder een aantal scans en hun automatische gegenereerd transcripties en de geconstateerde fouten en uitdagingen.

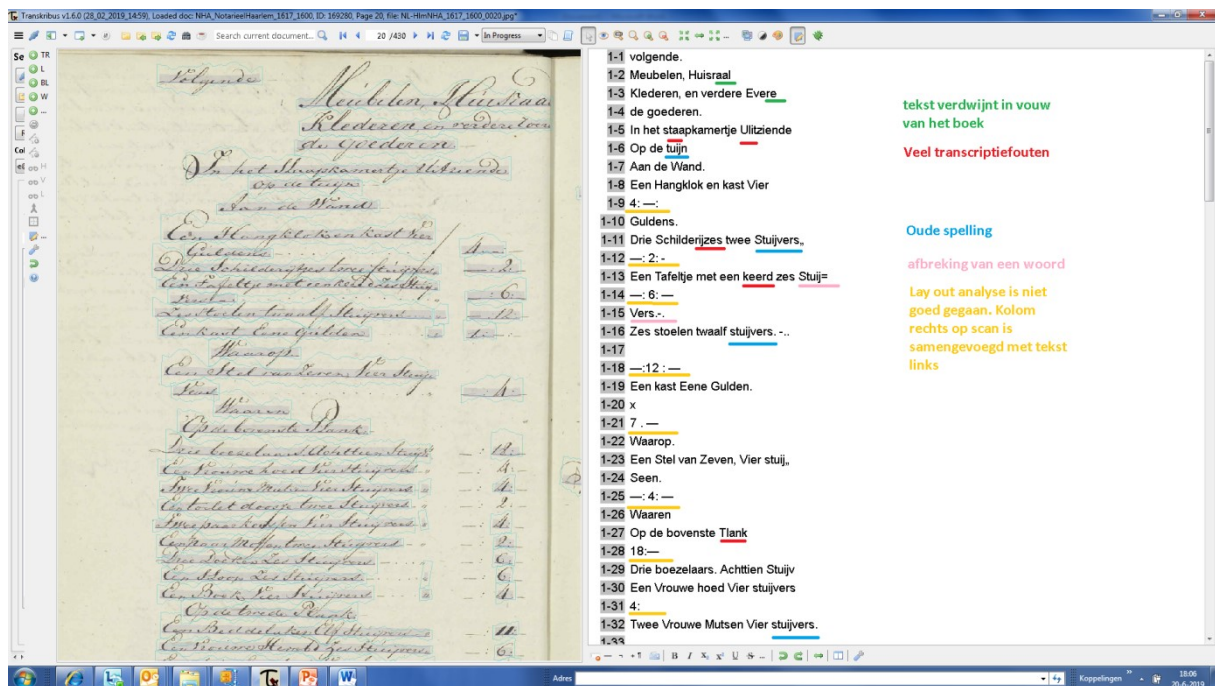




Afb 2) Scan van het archief van de VOC (toegangsnummer 1.04.02), inventarisnummer 7528, volgnummer 2010 en de transcriptie



Afb 3) Scan van het notarieel archief Haarlem (toegangsnummer 1617), inventarisnummer 1800, volgnummer 100 en de transcriptie



Afb 4) Scan van het notarieel archief Haarlem (toegangsnummer 1617), inventarisnummer 1600, volgnummer 20 en de transcriptie