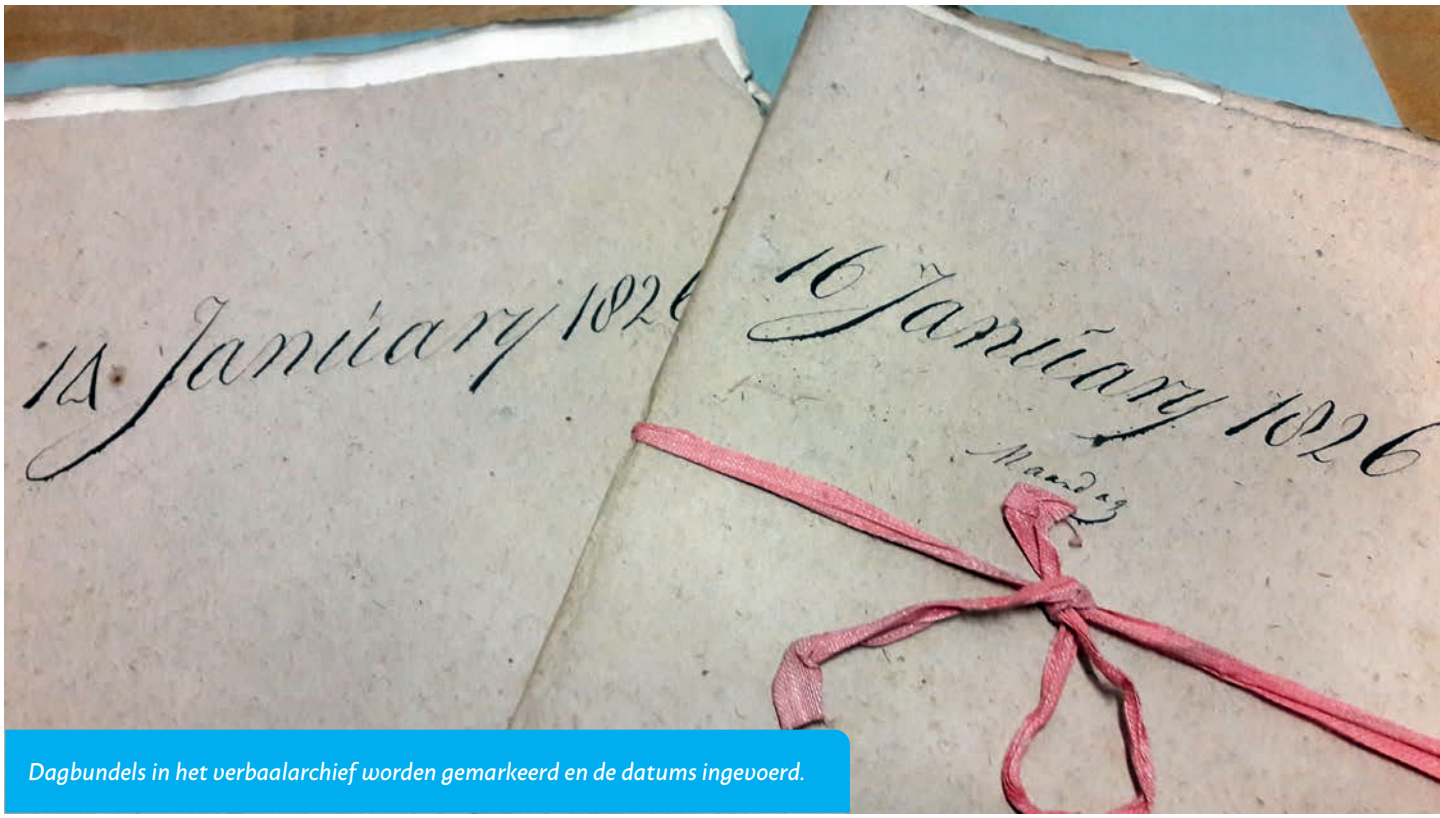
Tekst **Liesbeth Keijser** | projectleider Digitalisering bij het Nationaal Archief

Zoeken in scans wordt aantrekkelijker

Gedigitaliseerd archief beter toegankelijk

Het Nationaal Archief digitaliseert op grote schaal. Meters historische archieven zijn als plaatjes beschikbaar op internet. De toegankelijkheid van deze archieven wordt hiermee enorm vergroot. Ze zijn immers altijd en overal raadpleegbaar. Maar het bladeren door duizend scans van een archiefstuk van de VOC is minder aantrekkelijk dan het raadplegen van een in perkament gebonden deel waarbij je in één oogopslag een beeld hebt van de opbouw van een stuk in katernen en onderdelen.



Dagbundels in het verbaalarchief worden gemarkeerd en de datums ingevoerd.

Om de klant beter van dienst te zijn maken we de scans meer toegankelijk. Te digitaliseren archieven werden vóór het scannen al geïnspecteerd om de conserveringsbehoefte vast te stellen en om te bepalen wat de specificaties zijn voor digitalisering. Daar is nu een check aan toegevoegd om te weten hoe we de scans nader toegankelijk kunnen maken.

Navigeren in scans

Zo zijn we op dit moment bezig met een groot verbaal archief van het Ministerie van Koloniën uit de negentiende eeuw. Verbalen zijn besluiten die vaak zijn geordend in bundels per dag. Via een index kan de onderzoeker besluiten vinden over een bepaald onderwerp en zien op welke dag en met welk registratienummer een besluit is genomen. Fysiek zijn de dagbundels duidelijk herkenbaar. Door de eerste scan van een dagbundel te markeren en de vermelde datum op de omslag in te voeren met data-entry, kan de onderzoeker ook digitaal eenvoudig een dagbundel terugvinden. Binnen een dagbundel worden vervolgens de afzonderlijke archiefstukken (de agenda, besluiten en bijlagen) gemarkeerd. Deze zijn (nog) niet verwerkt met data-entry, maar door de markering wel beter zichtbaar.

Foto's **Nationaal Archief**

‘De resultaten zijn boven verwachting’

Het resultaat is een Excelbestand met een kolom met namen van gemarkeerde scans en een kolom met de datum van de dagbundel of de term ‘archiefstuk’. De website van het Nationaal Archief is nu nog in bewerking. Zodra dit gereed is, kan de klant met deze informatie eenvoudig navigeren naar de juiste dag binnen een inventarisnummer. Vervolgens kan hij ‘springen’ van archiefstuk naar archiefstuk om het juiste besluit te vinden.
Om een beeld te krijgen van wat de klant van deze verbeteringen vindt, heb ik met een collega bezoekers aan de studiezaal van het Nationaal Archief uitleg gegeven over onze plannen en feedback gevraagd. Alle ondervraagden waren enthousiast, hoewel een aantal het liefst ook data-entry wil op het niveau van het archiefstuk.

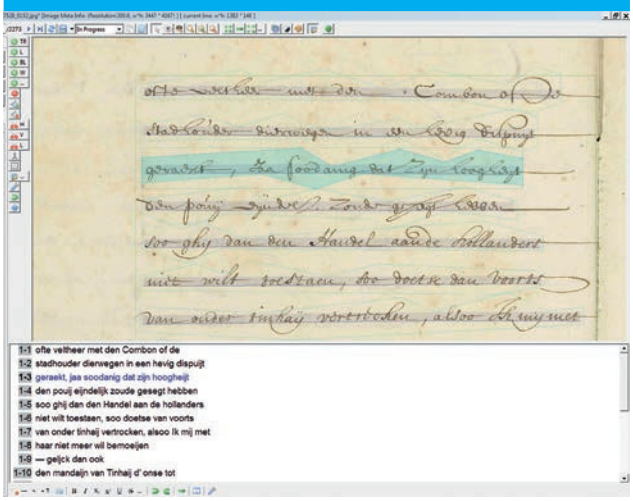
Leerpunt

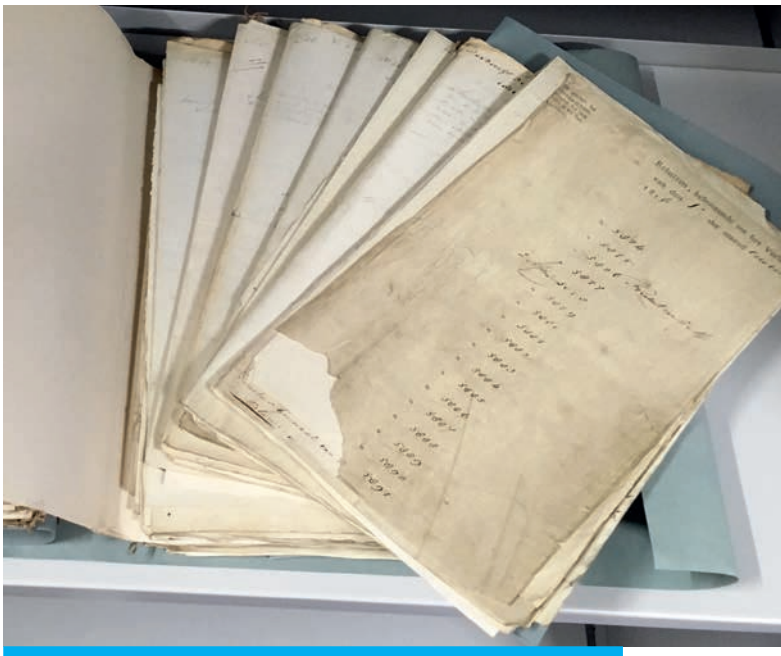
We gebruiken de fysieke verschijningsvorm van inventarisnummers om de scans te structureren en te presenteren. Bij een modern archief dat we laatst onderzochten, merkten we dat die structuur bij het herverpakken verloren was gegaan. Nietjes, paperclips, zure omslagen en multomappen waren verwijderd. Reconstructie van de opbouw van de onderdelen van een inventarisnummer wordt hierdoor aanzienlijk bemoeilijkt. Wellicht zou bij de conservering van archieven meer rekening gehouden moeten worden met het behoud van de structuur door nietjes, paperclips, zure omslagen en multomappen niet slechts te verwijderen, maar te vervangen door duurzame varianten.

Makkelijker zoeken

In archiefstukken komen we vaak eigentijdse indexen tegen, lijsten van zaken en personen met verwijzingen naar folio’s of pagina’s. Deze indexen zijn niet altijd vermeld in de beschrijving van het archiefstuk. Het te scannen archief

Scan van een VOC-stuk in Transkribus.





Archiefstukken in een dagbundel worden gemarkeerd.

wordt blaadje voor blaadje doorgenomen voor digitalisering. Tijdens dat proces markeren we deze indexen. Het zichtbaar maken van de locatie van deze indexen binnen een inventarisnummer op onze website kan het zoeken nog meer vereenvoudigen. En als een index van groot belang is, kunnen we deze met data-entry omzetten naar een digitale index. Bestaande digitale indexen kunnen worden verrijkt met de bijbehorende scans. Zo koppelden we bijna twee miljoen scans van de VOC aan de index van de VOC-opvarenden. Als een genealoog op zoek is naar een voorvader, dan vindt hij nu op de website van het Nationaal Archief niet alleen de naam, functie en informatie over zijn dienstreis, maar ook een scan met de originele registratie van zijn voorouder en een overzicht met een specificatie van de uitgaven die hij deed.

Automatische handschriftherkenning

Revolutionair is de mogelijkheid om ‘full text’ te kunnen zoeken in archieven. Dit gebeurt met ‘optical character recognition’ (OCR) bij getypte en gedrukte stukken. En bij handgeschreven documenten met automatische handschriftherkenning (handwritten text recognition/HTR). Het Nationaal Archief startte begin 2019 met het grootschalig transcriberen van scans met HTR in het platform Transkribus. We selecteerden hiervoor ongeveer één miljoen scans van het archief van de VOC uit de 17e en 18e eeuw van het Nationaal Archief en één miljoen scans van notariële archieven uit de 19e eeuw van het Noord-Hollands Archief en acht andere Regionaal Historische Centra.
Om de machine te leren automatisch te transcriberen, typte een transcriptieteam bijna 8000 pagina’s over in Transkribus. Met behulp van kunstmatige intelligentie ontwikkelden we hiermee trainingsmodellen, waaronder het model IJsberg. De resultaten zijn boven verwachting. Meer dan 90% van de tekens wordt correct herkend. We kunnen nu de twee »

» miljoen scans automatisch transcriberen. In Transkribus is het model IJsberg vrij beschikbaar, zodat iedereen het kan gebruiken om te transcriberen of om te gebruiken als basis voor een eigen model.

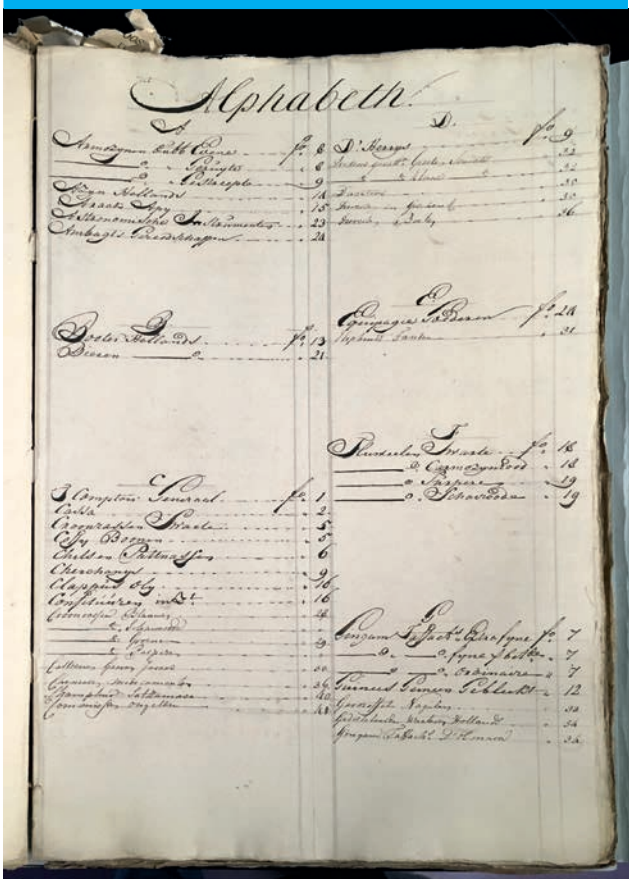
Zoeken in transcripties

Hiermee zijn we er nog niet. Ook de doorzoekbaarheid van deze transcripties willen we klantvriendelijk maken. Net als in Delpher voor het zoeken in kranten, willen we een scan en bijbehorende transcriptie gecombineerd tonen. En de zoekresultaten in beide 'highlighten'. Belangrijk is dat de onderzoeker in de grote hoeveelheid transcriptiedata door de bomen het bos nog ziet. De beschrijving van een archief kan helpen om te kunnen filteren op bijvoorbeeld jaartal van het archiefstuk of naam van de notaris. Maar we verwachten ook dat in de transcriptiedata bijvoorbeeld persoonsnamen en plaatsnamen kunnen worden gevonden, waarmee de zoekresultaten kunnen worden geduid. Het herkennen van deze namen heet Named Entity Recognition.
Begin volgend jaar is een projectwebsite met deze functionaliteit beschikbaar. Als de website van het Nationaal Archief er klaar voor is, willen we het daarin integreren. De gebruikte technologie is zo veel mogelijk open en herbruikbaar, zodat ook andere instellingen de functionaliteit kunnen nabouwen.

Vragen? Stel ze!

Hoewel niet veel archiefinstellingen in staat zijn om op grote schaal archieven te digitaliseren, hoop ik met dit artikel wat handreikingen te geven om een beter resultaat te halen uit digitaliseringsactiviteiten.
Heb je vragen naar aanleiding van dit artikel of ben je geïnteresseerd in ontwikkelingen op dat gebied? Op de website van het Nationaal Archief staat een artikel over automatische handschriftherkenning.[1] Je kunt je vragen ook stellen via het contactformulier op www.nationaalarchief.nl. |

Een index in een archiefstuk wordt gemarkeerd.



NOTEN

1 https://www.nationaalarchief.nl/archiveren/nieuws/succesvolle-resultaten-automatische-handschriftherkenning

VOC: Opvarenden

Periode: 1699 1794

Voornaam opvarende Jan
Patroniem opvarende Babtist
Tussenvoegsel opvarende de
Achternaam opvarende Zaan
Herkomst opvarende Oostkerke
Datum indiensttreding 1768-09-12
Functie bij indiensttreding Soldaat
Uitleg over functie Militair
Uitgevaren met het schip Oostkapelle
Datum uit dienst 1768-12-27
Waar uit dienst oostkapelle
Reden uit dienst Overleden
Schuldbrief Ja
Maandbrief Nee

Scan van de registratie van Jan Babtist, die is gekoppeld aan de index van de VOC-opvaarenden.