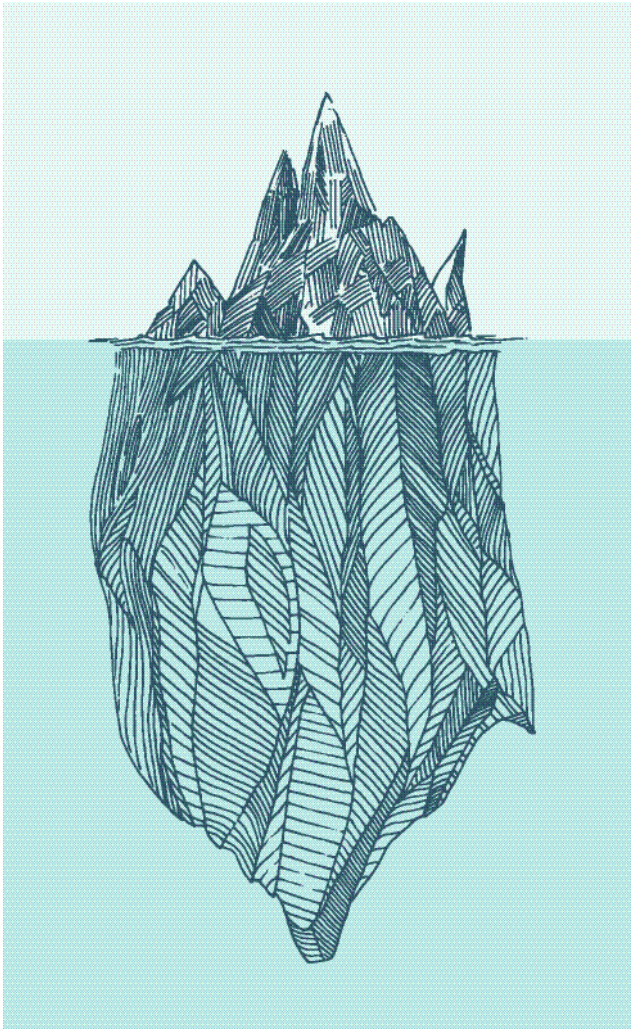
Projectplan
De ijsberg zichtbaar maken

Versie	5.1
Datum	15 november 2018
Status	definitief
Auteur	Liesbeth Keijser
Instelling	Nationaal Archief



Het topje van de ijsberg, dat zijn de beschrijvingen die we hebben van archieven. Maar het grootste deel van de ijsberg is de informatie die in al die documenten zelf staat. Deze informatie is vaak wel gedigitaliseerd maar is niet machine leesbaar. Die informatie willen we boven water halen met handschriftherkenning.

Aanleiding

In het kader van intensivering nationale strategie digitaal erfgoed/NDE heeft het Ministerie van OCW, directie M&C, middelen beschikbaar voor handschriftherkenning (zie EDOC-#1307849-v2-Fiche_RA_middelen_nationale_strategie_digitaal_erfgoed). Het Nationaal Archief is gevraagd hiervoor een projectplan te maken. Als mogelijke betrokkenen zijn genoemd het Huygens Instituut voor Nederlandse Geschiedenis en de regionale archieven.

De algemene doelen van de extra middelden in het kader van intensivering nationale strategie digitaal erfgoed/NDE zijn 'duurzame en generieke aanpak van digitalisering cultureel erfgoed' en 'vergroten bruikbaarheid van digitale erfgoedcollecties.'

Doel van het project

Erfgoedinstellingen hebben de laatste jaren veel gedigitaliseerd. Historische handschriftelijke collecties en archieven zijn als plaatjes beschikbaar op internet. Maar wat veel mensen zich niet realiseren is dat de informatie in deze documenten niet doorzoekbaar is. Scans zijn niet machine leesbaar en zijn digitaal dus niet vindbaar voor onderzoekers, studenten, journalisten of grasduiners. Handschriftherkenning gaat hier verandering in brengen.

Met het innovatieve project *De ijsberg zichtbaar maken* gaan we aan de slag met handschriftherkenning. Grote datasets van het Nationaal Archief en de Regionaal Historische Centra ontsluiten we met behulp van het semi-automatisch transcriptie platform Transkribus. Na een training herkent de software het handschrift en de taal. De scans van handschriften worden vervolgens automatisch getranscribeerd. Om de gebruiker goed te faciliteren ontwikkelen we een functionaliteit voor het zoeken in de transcripties en het tonen van de resultaten. Voor het eerst in de geschiedenis kunnen we automatisch door handgeschreven archieven zoeken.

Iedereen kan profiteren van de getrainde transcriptie software. De geproduceerde transcripties komen beschikbaar als Open Data. Uitgangspunt is dat de ontwikkelde functionaliteiten voor zoeken en tonen van de transcripties herbruikbaar en Open Source zijn. De opgedane kennis wordt gedeeld met het veld.

Transkribus

Achtergrond

We maken gebruik van het transcriptieplatform Transkribus. Dit platform werd opgericht door het project READ (Recognition and Enrichment of Archival Documents 2016-2019), een Europese H2020 e-Infrastructuur project, onder leiding van de Universiteit van Innsbruck met de medewerking van onderzoeksgroepen uit heel Europa (http://read.transkribus.eu/).

We kiezen voor Transkribus omdat het een vrij toegankelijk platform is waarin de gehele transcriptie workflow kan worden ingericht. Het heeft 13.000 gebruikers en 70 samenwerkingsovereenkomsten met instellingen in meer dan 20 landen. Er is een 'Dutch model' in het platform zodat instellingen kunnen profiteren van elkaars Nederlandse transcripties. Al veel Nederlandse data zijn beschikbaar. Diverse culturele instellingen in Nederland gaan op korte termijn van start met grote projecten in Transkribus. Voor ons project zal een Open Source handschrift herkenningspakket worden gebruikt waardoor na de projectperiode het automatisch transcriberen kan worden voortgezet zonder licentiekosten.

In 2018 startte het National Archief samen met het Noord-Hollands Archief een pilot met Transkribus. Vrijwilligers transcriberen een beperkte hoeveelheid scans van archieven van de twee instellingen. Hiermee wordt de machine getraind op deze handschriften en kan worden getest wat het resultaat is van de training. De eerste testresultaten zijn bemoedigend. Met een beperkte trainingset wordt een accuraatheid van het aantal automatische getranscribeerde karakters bereikt van 80,24%, waarbij een groot deel van de fouten misinterpretaties zijn van punten en komma's.

Coöperatie

Het project READ eindigt medio 2019. Om de onderzoeksinfrastructuur en het opgebouwde netwerk te consolideren en te kunnen uitbreiden wordt een European Cooperative Society (SCE) opgericht die zetelt in de Universiteit van Innsbruck, (een van) de oprichters. De SCE wordt gevormd door leden, tevens aandeelhouders. Uit hun midden worden een Raad van Bestuur en een Raad van Commissarissen gekozen. De coöperatie verricht diensten tegen betaling. Het Nationaal Archief zal geen lid worden van de SCE maar middels samenwerkingsafspraken aan de SCE worden verbonden. We onderzoeken welke vorm recht doet aan de samenwerking en niet tot juridische bezwaren leidt.

Activiteitenplan

Kennisdelen

Bij het schrijven van het projectplan werd het Nationaal Archief geconfronteerd met allerlei vragen. Bijvoorbeeld: Waar moeten de handschriften van een dataset aan voldoen? Wordt een automatische transcriptie beter als een lexicon wordt toegevoegd aan het trainingsmodel? Kunnen we het resultaat verbeteren met nabewerkingen. Waar moeten de transcripties worden opgeslagen? Hoe willen we de transcripties doorzoekbaar maken en tonen.

Tijdens het project wordt een groot deel van deze vragen beantwoord. Het veld wordt uitgenodigd mee te denken. Hiervoor gebruiken we het platform Kennisnetwerk Informatie en Archief (KIA). Onze ervaringen met automatische handschriftherkenning delen we o.a. via dit medium.

Transcriberen

Het Nationaal Archief heeft de afgelopen jaren bijna 4 kilometerarchief gedigitaliseerd van zowel haar eigen collectie als die van de Regionaal Historische Centra. De instellingen selecteerden hiervoor hun belangrijkste archiefstukken. Het Nationaal Archief digitaliseerde onder andere het gehele archief van de VOC. De Regionaal Historische Centra kozen het notarieel archief uit de 19e eeuw, een belangrijke bron voor genealogisch onderzoek. In overleg met de Regionaal Historische Centra worden hieruit scans gedestilleerd voor het trainen en testen van het platform Transkribus en voor het automatische verwerken van de scans na de training. Mogelijke factoren die meespelen bij de selectie zijn de leesbaarheid van het schrift en de uniformiteit van de lay-out en het handschrift. Wellicht zijn de handschriften van Willem van Oranje of Johan de Wit goed bruikbaar voor het project.

Door het handmatig transcriberen van de scans in het platform wordt de machine door de mens getraind. De transcripties worden gemaakt door een leverancier van Transkribus. Iemand getraind in het lezen van oude handschriften controleert en corrigeert de transcripties.

Voor de verschillende datasets worden trainingsmodellen gemaakt door Transkribus in samenwerking met het NA. We onderzoeken hoe we het beste resultaat bereiken. Zowel de kwaliteit van de transcripties als de resultaten van de trainingsmodellen worden geëvalueerd.

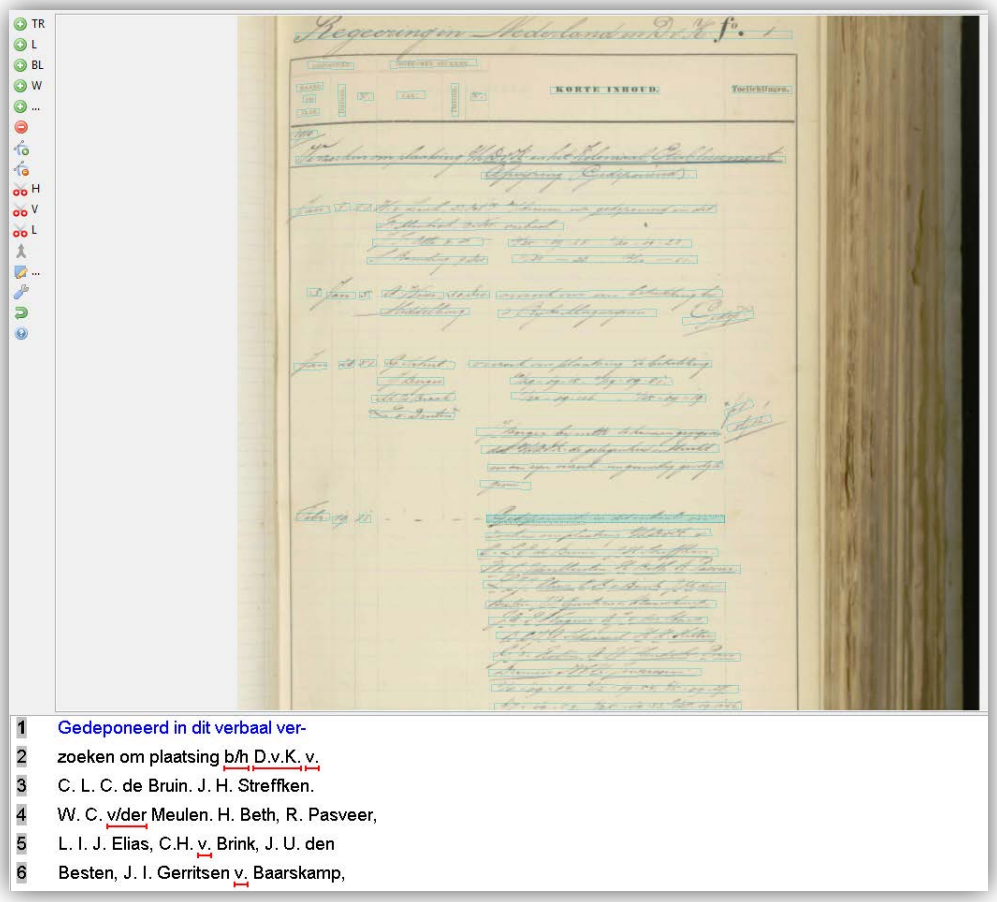
Het Noord-Hollands Archief en het Nationaal Archief hebben hiermee al ervaring opgedaan in het lopende pilot project. In het project *De ijsberg zichtbaar maken* treden we hier gezamenlijk in op.

Als de resultaten van de trainingsmodellen voldoende zijn, kunnen vervolgens scans automatisch worden getranscribeerd door Transkribus. We streven hierbij naar een accuraatheid van de herkenning van de karakters van minimaal 80%.

We gaan 2 miljoen scans verwerken in Transkribus. Hiermee ontsluiten we circa 9% van de huidige DTR productie. Met de ontwikkelde trainingsmodellen kunnen na het project nog meer scans toegankelijk worden gemaakt. Het percentage ontsloten archieven zal dan fors toenemen.

Ontsluiten bestaande transcripties

Een aantal belangrijke handschriftelijke documenten van het Nationaal Archief is al getranscribeerd voor tentoonstellingen en publicaties. De dagregisters van Kaap de Goede Hoop van het archief van de VOC zijn getranscribeerd door het Zuid-Afrikaanse Tracing History Trust in samenwerking met anderen. We onderzoeken of we met hen kunnen samenwerken. Om ook deze transcripties te kunnen doorzoeken en tonen, gaan we ze in Transkribus koppelen aan hun scans. De transcripties worden bovendien ook gebuikt om het transcriptieplatform te trainen.



Transcriberen in het platform Transkribus.

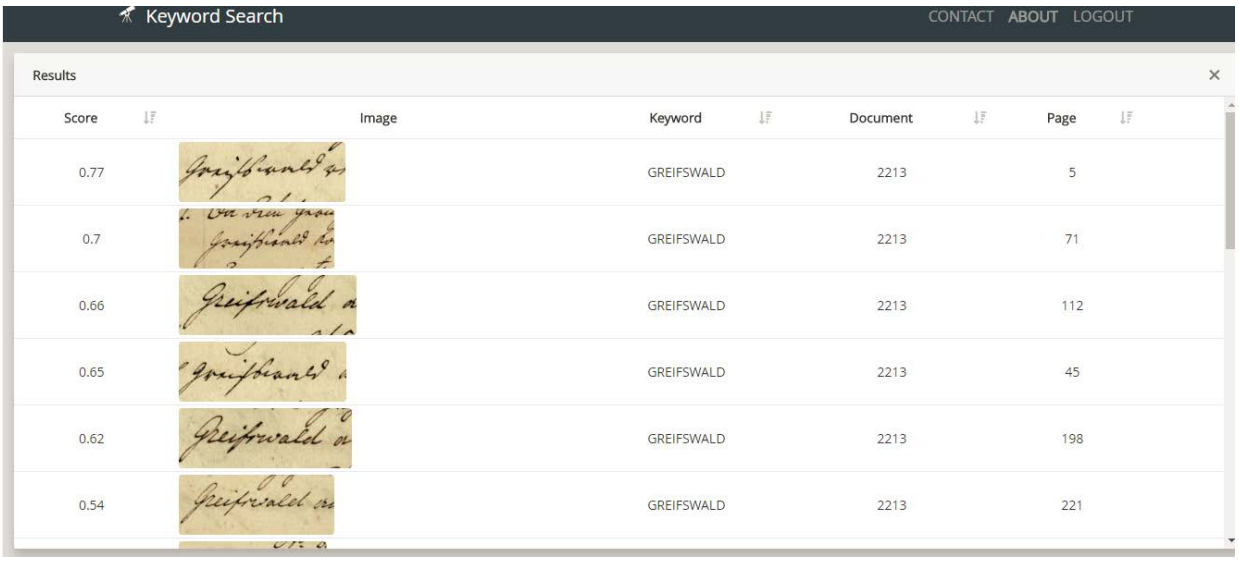
Nabewerking van de transcriptieresultaten

Met nabewerking kunnen relevante gegevens uit de verkregen data worden gedestilleerd. Zo kunnen mogelijk personen, organisaties en locaties worden getagt of kan het onderwerp van een archief automatisch worden bepaald.

Zoeken en tonen van de transcripties

Er wordt een functionaliteit ontwikkeld voor het zoeken in de transcripties en voor het tonen van de zoekresultaten. De infrastructuur van het Nationaal Archief en de Regionaal Historische Centra is nog in ontwikkeling. Daarom wordt gestart met een onderzoek door een solution architect naar de mogelijkheden van een herbruikbare en Open Source oplossing, die in dien mogelijk aansluit bij de infrastructuur van het Nationaal Archief en de Regionaal Historische Centra. Bij voorkeur zijn de functionaliteiten ook geschikt voor het zoeken en tonen van resultaten van Optical Character Recognition van getypt en gedrukt archiefmateriaal. De solution architect onderzoekt eveneens hoe de transcripties (duurzaam) moeten worden opgeslagen en beheerd.

We willen voor de zoekfunctionaliteit gebruik maken van de techniek Keyword Spotting. Een techniek waarbij de zoekresultaten worden getoond met een score voor confidentie en relevantie. Een zoekterm kan hierdoor ook worden gevonden als die niet helemaal correct is getranscribeerd of als er sprake is van een spellingsvariant.



Prototype web interface voor Keyword Spotting met score voor confidentie en relevantie

De functionaliteit voor het tonen van de zoekresultaten, toont de scan met daarop het zoekresultaat gehighlight. De zoekresultaten kunnen worden gefilterd op bijvoorbeeld archief, inventarisnummer, plaats, tijd en persoon. In geval van een gevalideerde transcriptie is het mogelijk de transcriptie in zijn geheel te tonen.

De gebruiker staat centraal in dit project. Een gebruikersonderzoek is onderdeel van de ontwikkeling van deze functionaliteit.

Overzicht resultaten

Korte termijn (projectperiode):

- Met dit project bouwen we kennis op van semi-automatische handschriftherkenning. Een innovatieve techniek die scans van handgeschreven documenten machine leesbaar maakt.
- Om het erfgoedveld te laten profiteren van deze kennis plaatsen we regelmatig berichten op KIA, een platform van het Kennisnetwerk Informatie en Archief.
- In het platform Transkribus worden trainingsmodellen gemaakt voor de geselecteerde datasets. Hiermee worden scans automatisch getranscribeerd.
- Onderdeel van die trainingsmodellen zijn de door de mens gemaakte transcripties. We gaan 6000 scans laten transcriberen door de mens.
- Om bestaande transcripties te kunnen doorzoeken en tonen, gaan we ze in Transkribus koppelen aan hun 5000 scans. Deze zullen worden toegevoegd aan de trainingsmodellen.
- Met de trainingsmodellen gaan we 2 miljoen scans automatisch transcriberen.
- Er wordt een onderzoek gedaan naar een herbruikbare en Open Source functionaliteit voor het zoeken en tonen van de transcripties
- Er komt een herbruikbare en Open Source functionaliteit beschikbaar voor het zoeken in en tonen van de transcripties.
- Met dit project laten we zien dat OCW en het Nederlandse archiefveld op innovatieve wijze cultureel erfgoed toegankelijk maken.

Lange termijn (na de projectperiode)

- We dragen bij aan het beschikbaar houden van het semi-automatische handschriftherkenningsplatform Transkribus.
- De ontwikkelde trainingsmodellen zijn beschikbaar voor iedereen binnen Transkribus. Hiermee kunnen na het project nog meer scans toegankelijk worden gemaakt.
  De door de mens gemaakte transcripties zijn beschikbaar voor iedereen. Deze data kan worden gebruikt voor trainingsmodellen voor andere collecties in Transkribus en andere handschriftherkenningssoftware.

Scope

Binnen de scope van het project vallen:

- De in het activiteitenplan genoemde activiteiten.
- Het automatisch vervaardigen van transcripties waarbij we streven naar een accuraatheid van de herkenning van de karakters van minimaal 80%.
- Het uitzoeken waar de transcripties opgeslagen moeten worden.
- Onderzoeken of kan worden voorgesorteerd op Linked Open Data.

Buiten de scope van het project vallen:

- Het digitaliseren van archieven.
- Het verwerken van archieven met een complexe lay-out.
- Het vervaardigen van 100% correcte transcripties door de machine.
- Het vervaardigen van een functionaliteit voor het zoeken en tonen van Linked Open Data.
- Het beheer van de transcripties.

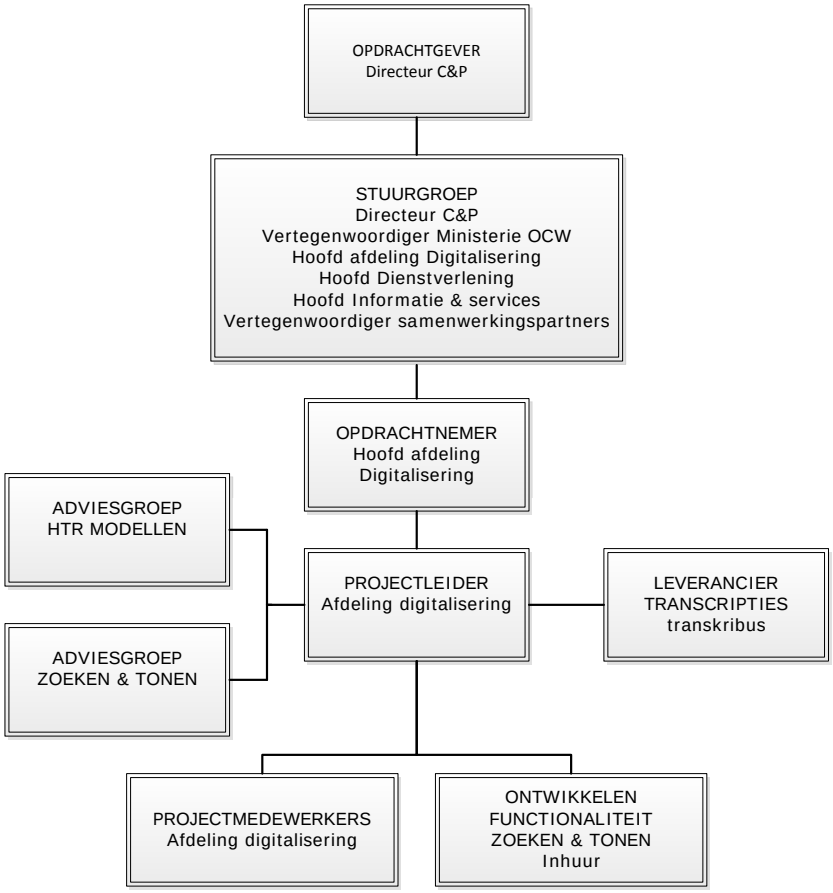
Organisatie

De projectorganisatie is weergegeven in onderstaand schema. Opdrachtgever is de Directeur Collectie & Publiek van het Nationaal Archief.

Circa één maal in de 2 maanden komt de stuurgroep bijeen. Voorzitter is de Directeur Collectie & Publiek. Op de agenda staan onder ander de voortgang van het project, de mijlpalen, de financiën, de knelpunten en de risico's.

De adviesgroepen worden gevormd door belanghebbenden en inhoudelijk deskundigen van onder andere de samenwerkingspartners. De vergaderfrequentie is afhankelijk van het stadium van het project.

Door inhuur van capaciteit voor het ontwikkelen van een functionaliteit voor zoeken en tonen wordt de belasting van de afdeling Informatie & Services zoveel mogelijk beperkt. Om voldoende aansluiting te houden met de ontwikkeling van de infrastructuur van het Nationaal Archief en de RHC's wordt regelmatig overlegd met vertegenwoordigers van deze afdeling.



Samenwerken

Het project wordt gerealiseerd in nauwe samenwerking met het project READ en zijn opvolger, een European Cooperative Society.

Het Noord-Hollands Archief en het Nationaal Archief treden gezamenlijk op in de evaluatie van de kwaliteit van de transcripties en de resultaten van de trainingsmodellen.

Het Nationaal Archief en de Regionaal Historische Centra selecteren de scans voor het trainen en testen van het platform en voor het automatische verwerken van de scans na de training.

Huygens Instituut voor Nederlandse Geschiedenis wordt in het project geconsulteerd als gebruiker van de archieven. De wederzijdse ervaringen met Transkribus worden gedeeld om het resultaat te verbeteren.

Planning

ACTIVITEITEN	2019				2020			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Kennisdelen								
Een maal per kwartaal een bericht op KIA.								
Transcriberen								
Selectie scans.								
Uploaden scans naar Transkribus.								
Transcriberen door mens t.b.v. inleren machine.								
Controle transcripties.								
Maken van trainingsmodellen.								
Transcriberen door machine.								
Ontsluiten bestaande transcripties								
Koppelen scans aan bestaande transcriptie.								
Zoeken en tonen van de transcripties								
Onderzoek naar herbruikbare en Open Source functionaliteit.								
Ontwikkelen functionaliteit zoeken in transcripties.								
Ontwikkelen functionaliteit tonen zoekresultaten transcripties.								

Uitgangspunten

De volgende uitgangspunten zijn van toepassing voor dit project:

- We maken gebruik van het handschriftherkenning platform Transkribus.
- Te ontwikkelen functionaliteiten voor zoeken in en tonen van de transcripties passen in de digitale infrastructuur van Nationaal Archief en Regionaal Historische Centra.
- Te ontwikkelen functionaliteiten voor zoeken en tonen van de transcripties zijn herbruikbaar en Open Source.
- De trainingsmodellen zijn voor iedereen beschikbaar in Transkribus.
- De transcripties zijn beschikbaar als Open Data.
- Er wordt een solution architect ingehuurd. Deze wordt ingezet op het project of is ter compensatie van een interne kandidaat.
- Er worden ontwikkelaars ingehuurd. Deze worden ingezet op het project of zijn ter compensatie van interne kandidaten.