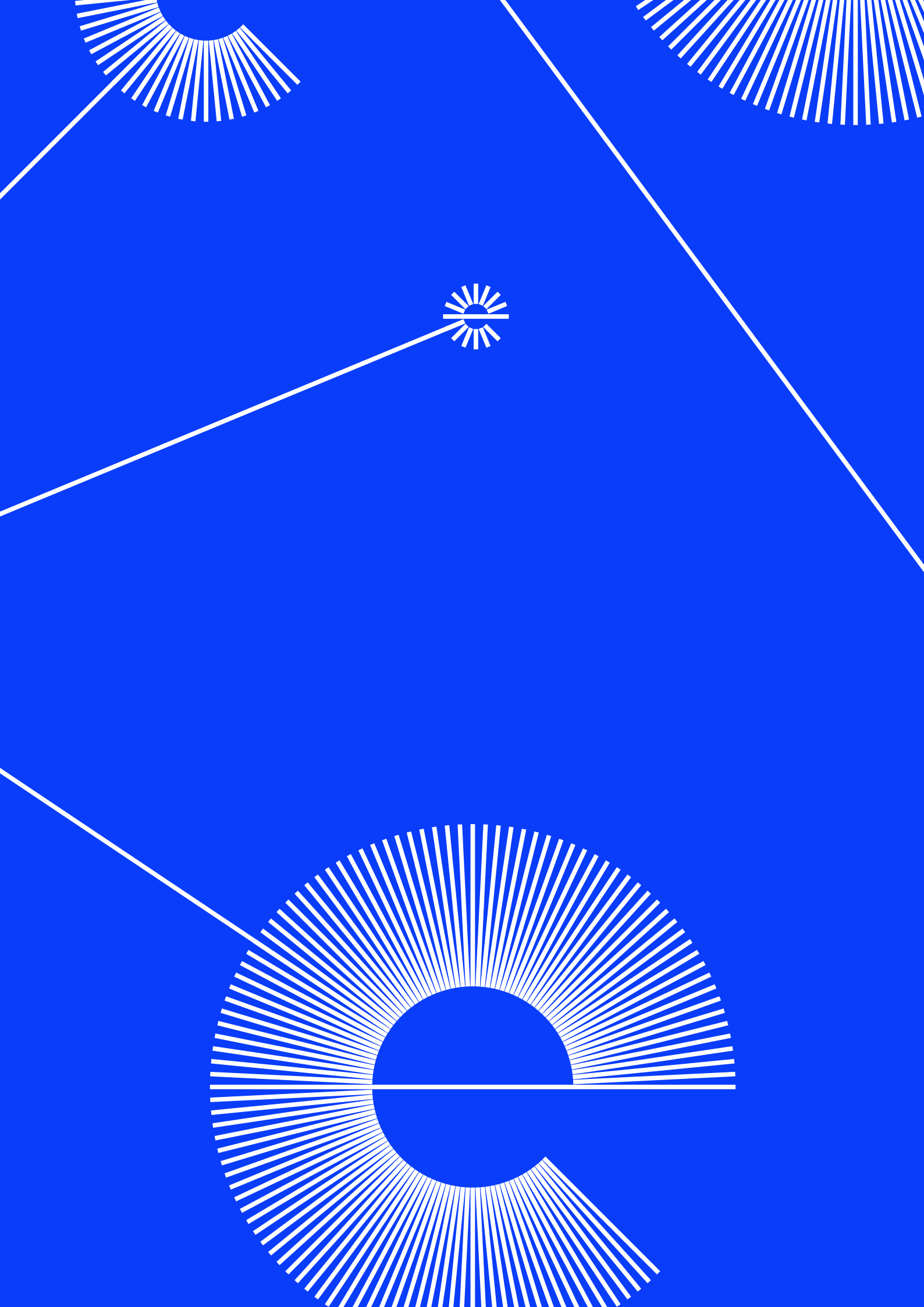
netwerk digitaal erfgoed

Krant & foto's verbonden

Een verkenning om kunstmatige intelligentie in te zetten om erfgoedcollecties te verbinden



Inhoudsopgave

Inleiding 4

AI in erfgoed 4

Collecties 5

Aanpak 5

Hoofdstuk 1 - De collecties met persfoto's en krantenfoto's 6

Collecties als data 6

Juridische aspecten bij het gebruik van kranten en foto's 7

Hoofdstuk 2 - Slimme algoritmiek om krantenfoto's en persfoto's te matchen 8

In stappen naar een model voor beeldherkenning 8

De resultaten 10

Juridische aspecten bij het toepassen van beeldherkenning 10

Hoofdstuk 3 - Het zoeken door de foto's: de demo en de gebruiker 11

De demonstrator: een interface als proof of concept 11

De reportage achter de persfoto 11

Matches door de computer én de gebruiker 12

Keuzes in de presentatie in de demo 13

Gebruiksonderzoek 13

Juridische aspecten bij het ontwikkelen van een interface 15

Hoofdstuk 4 - Resultaten en inzichten: een samenvatting 16

De resultaten 16

Inzichten 16

Epiloog: Uitdagingen voor de toekomst 18

Verbeteringen van de resultaten binnen Krant en Foto's 18

Frisse ideeën voor de toekomst 19

Colofon 20

Inleiding

Foto's vertellen verhalen en staan nooit op zichzelf. Ze zijn een momentopname van de werkelijkheid. Ook in kranten worden foto's gebruikt om een verhaal te vertellen. Maar waar komen die foto's vandaan? Vaak blijkt het ingewikkeld om die geschiedenis te achterhalen. Kranten en fotoarchieven worden al lange tijd gedigitaliseerd en ontsloten, maar de link tussen beide ontbreekt nog. Zo zijn veel persfotoarchieven gedigitaliseerd en beschikbaar bij regionale archieven, terwijl veel kranten digitaal in te zien zijn bij online platforms, zoals Delpher. Daarom is het vooralsnog alleen mogelijk om foto's te zoeken in aparte beeldbanken door de vooraf beschreven tekstuele kenmerken van de foto in de metadata te gebruiken.

De opkomst van kunstmatige intelligentie (AI) maakt het nu mogelijk om beeldherkenning in te zetten voor collecties uit de cultuursector. Hierdoor kunnen we niet alleen op de formele beschrijving van een foto zoeken, maar ook foto's ontdekken op basis van de beeldkenmerken. Daarnaast kunnen we verbindingen leggen tussen de foto in de kranten met de bijbehorende beschrijving in het artikel en de oorspronkelijke serie foto's in het fotoarchief. Dit maakt het mogelijk voor elke geïnteresseerde om foto's in hun context te bestuderen en het verhaal achter een foto te leren kennen.

In dit whitepaper beschrijven we de resultaten en lessen die we hebben opgedaan bij het onderzoeksproject *Krant en Foto's*, in de periode juli 2021 tot maart 2022. Hiermee hopen we collega-erfgoedinstellingen te inspireren om ook actief aan de slag te gaan met kunstmatige intelligentie om hun digitale collecties op nieuwe manieren te ontsluiten en verbinden. Op die manier zijn meer mensen in staat om op meer manieren digitale erfgoedschatten te ontdekken, verkennen en gebruiken.

AI in erfgoed

Binnen de erfgoedwereld wordt steeds meer geëxperimenteerd met de inzet van AI in het algemeen en beeldherkenning in het bijzonder. Erfgoed Gelderland heeft bijvoorbeeld al eens de beeldherkenningstechniek van Google ingezet om afbeeldingen te voorzien van metadata.[1] In Vlaanderen wordt eenzelfde soort toepassing verkend door gezichtsherkenning in te zetten om personen op foto's en video's te herkennen.[2] In het Verenigd Koninkrijk verkent onderzoeker Giles Bergel met de Oxford Visual Geometry Group al jarenlang de mogelijkheden van AI om beeldcollecties op nieuwe manieren te ontsluiten.[3]

Krant en Foto's bouwt voort op deze experimenten en sluit aan bij de ambitie die geformuleerd is in de Nationale Strategie Digitaal Erfgoed om met behulp van AI en big data-technieken 'bredere, meerstemmige en participatieve manieren van collectieontsluiting' op te zetten.[4] Het project is tevens een use case van de werkgroep Cultuur en Media van de Nederlandse AI Coalitie.[5]

1 Netwerk Digitaal Erfgoed. Erfgoed Gelderland ging aan de slag met beeldherkenning. Dit zijn de geleerde AI-lessen. Geraadpleegd op 16-2-2022.
2 Meemoo. FAME: gezichtsherkenning als tool voor metadatacreatie. Geraadpleegd op 16-2-2022.
3 Bodleian Ballads. http://balladsblog.bodleian.ox.ac.uk/blog/570 & https://www.robots.ox.ac.uk/~vgg/research/. Geraadpleegd op 16-2-2022.
4 Netwerk Digitaal Erfgoed & ministerie van Onderwijs, Cultuur en Wetenschap. Nationale Strategie Digitaal Erfgoed. Den Haag: Netwerk Digitaal Erfgoed & ministerie van Onderwijs, Cultuur en Wetenschap, 2021.
5 Nederlandse AI Coalitie. Toepassingsgebieden. Geraadpleegd op 16-2-2022.

Collecties

In dit project hebben we de fotocollectie van Fotopersbureau De Boer (Noord-Hollands Archief) en Persfotobureau D. van der Veen (Groninger Archieven) verbonden met de publicatie van de foto's in *Haarlems Dagblad*, *IJmuider Courant* en *Nieuwsblad van het Noorden*. Hiertoe hebben we een demonstrator gemaakt waarmee het mogelijk is om een persfoto te vinden en de bijbehorende publicatie in de krant te bekijken. Daarnaast is het een optie om feedback te geven op de resultaten: kloppen de getoonde kranten of niet?

Aanpak

We hebben het project stapsgewijs uitgevoerd. Als eerste hebben we bepaald met welke datasets we gingen werken en deze met elkaar uitgewisseld. Vervolgens hebben we het algoritme en de interface ontwikkeld waarmee gebruikers de foto's kunnen verkennen. Dit whitepaper is opgebouwd volgens deze stappen. In elk hoofdstuk beschrijven we de stappen die we genomen hebben, de lessen die we daarbij geleerd hebben en de juridische perspectieven die per stap hebben meegespeeld. Tot slot schetsen we een beeld van de meerwaarde van de resultaten en lessen uit dit project voor de erfgoedsector en van de uitdagingen die voor ons liggen.

Processtap	Producent	Toelichting
Data verzamelen en aanleveren	Picturae, NHA, GrA, KB	Beeldbestanden van persfoto's en krantenfoto's en metadata. Een deel was al aanwezig in het collectiebeheersysteem van Picturae
Voorbereiding trainingsdata	NHA, GrA	Handmatig aanbrengen van verbindingen tussen kranten- en persfoto's als basis voor ontwikkeling van een model voor beeldherkenning
Ontwikkelen algoritme	Sioux	Algoritmes verkennen, testen en bijstellen. Maatstaf kwaliteit vaststellen, foto's uit kranten isoleren
Toepassen algoritme op grote dataset	Sioux	Definiëren outputformaat, omgeving opzetten voor opschalen, beeldbestanden persfoto's pre-processen, pipeline bouwen voor verwerken grote dataset, resultaten analyseren
Opleveren algoritme	Sioux	Robuuste code maken voor herbruikbaar proces, documentatie code schijven, overdracht voor Picturae schrijven, inclusief overzicht met alle gemaakte verbindingen
Ontwikkelen demo	Picturae	Interface met beeldbank en zoekfunctionaliteit op basis van datasets en verbindingen
Gebruiksonderzoek	NHA, GrA	Vragenlijsten opstellen, feedback van de gebruikers over de demo, interpreteren en vastleggen van resultaten
Release van resultaten	Picturae, Sioux, NHA	Demo, software voor beeldherkenning, uitkomst gebruiksonderzoek
Kennisdeling	Allen	Whitepaper met lessen en inzichten dient als basis voor allerlei andere vormen van kennisdeling, zoals artikelen, presentaties en videoclip

1. De collecties met persfoto's en krantenfoto's

Collecties als data

De persfotocollecties
In totaal omvatten de collecties van Fotopersbureau De Boer (1945-2004) en Persfotobureau D. van der Veen (1963-1980) respectievelijk zo'n twee miljoen en vierhonderdduizend persfoto's. We hebben ons gericht op de foto's en kranten uit de jaren zeventig, omdat beide in die periode overlappen en omdat de hoeveelheid bestanden binnen de projectperiode te verwerken leek.

Er zit een verschil in de aangeleverde datasets. De afbeeldingen van De Boer zijn losgeknipt, die van Van der Veen zijn complete negatiefstroken. Van deze negatiefstroken zijn losse afbeeldingen gemaakt. Aangezien deze losse afbeeldingen van Van der Veen slechts dienen als invoer voor de AI en niet worden opgeslagen en dus niet getoond in de demonstrator, heeft dit effect op de beeldbank.[6]

Kinderen spelen rolschaatshockey op de Parkweg in Groningen tijdens een autoloze zondag (1973). Negatiefstrook, weergegeven als contactafdruk. Foto's: Persfotobureau D. van der Veen. Collectie: Groninger Archieven. CCo.

De uitgebreide en goed op elkaar aansluitende metadata van de persfoto's zijn al beschikbaar vanuit Groninger Archieven en Noord-Hollands Archief. Ze zijn niet ingezet om de koppeling tussen foto's tot stand te brengen.

De krantencollecties
Van *Nieuwsblad van het Noorden* zijn de scans gebruikt van kranten die op Delpher toegankelijk zijn.[7] *Haarlems Dagblad* en *IJmuider Courant* zijn voor de jaren zeventig nog niet online beschikbaar, maar scans waren voor het project al wel te gebruiken.

6 Hoofdstuk 3 gaat in op de beeldbank.
7 Delpher is de nationale infrastructuur voor toegang tot volledige teksten met momenteel 130 miljoen pagina's aan boeken, kranten, tijdschriften en ANP-radiobulletins. Materiaal van fotografen of auteurs die een opt-out hebben gestuurd is niet toegankelijk via Delpher en niet gebruikt in dit project.

Dataset (periode 1970-1979)[8]	Bron	Aantal fotobestanden
Fotopersbureau De Boer	noord-hollands archief	410.879
Persfotobureau D. van der Veen	GRONINGER ARCHIEVEN	90.597
Haarlems Dagblad	noord-hollands archief	90.732
IJmuider Courant	noord-hollands archief	76.044
Nieuwsblad van het Noorden	GRONINGER ARCHIEVEN	95.620

Juridische aspecten bij het gebruik van kranten en foto's

Auteursrecht op krantenartikelen en foto's
De gekozen kranten zijn gepubliceerd in de periode 1970-1979 en daarom nog auteursrechtelijk beschermd. De rechten van de uitgevers vervallen immers zeventig jaar na publicatie. De rechten van freelancers, waaronder veel fotografen, vervallen zeventig jaar na het overlijden van de maker. Zowel bij het uitwisselen van data als de beschikbaarstelling van de thumbnails op de demo-website en de beschikbaarstelling van de kranten op de website(s) van de instellingen moet rekening gehouden worden met auteursrecht.

Voor het beschikbaar stellen van (mogelijk) auteursrechtelijk beschermde kranten op hun websites maken collectiehoudende instellingen afspraken met uitgevers en de collectieve beheersorganisaties (CBO's) Stichting Lira voor tekst en Pictoright voor beeld. Volgens deze afspraken mogen auteursrechtelijk beschermde werken niet zonder beperkingen doorgeleverd worden. Daarom zijn voor dit project afspraken gemaakt met Picturae en Sioux Technologies, waarin zij verklaren de geleverde data alleen voor dit project te gebruiken en daarna weer te verwijderen.

Persoonsgegevens in krantenartikelen en foto's
Naast het auteursrecht dient rekening gehouden te worden met het privacyrecht. De foto's in de kranten kunnen persoonsgegevens bevatten, namelijk als het foto's zijn van herkenbare, levende mensen. Voor de doorlevering van de foto's aan leveranciers Picturae en/of Sioux moeten daarom verwerkersovereenkomsten worden afgesloten. Dit moeten de instellingen zelfstandig doen wanneer zij persoonsgegevens doorleveren aan de leverancier(s), omdat zij daar als zelfstandig verwerkingsverantwoordelijke toe verplicht zijn volgens de Algemene Verordening Gegevensbescherming (AVG).

Auteursrecht persfotoarchieven
De auteursrechten van de collectie van Fotopersbureau De Boer en Persfotobureau D. van der Veen berusten bij respectievelijk Noord-Hollands Archief en Groninger Archieven. De collecties waren al voor hergebruik beschikbaar gesteld met een publieke domein-licentie 'CC0'.[9] Voor de uitvoering van het project *Krant en Foto's* hebben de rechten voor deze collecties dan ook geen rol gespeeld.

8 Bij de persfoto's van De Boer gaat het om losse foto's, bij Van der Veen om aantal negatiefstroken (met elke zo'n 2-4 foto's), bij de drie kranten om scans van de gehele pagina.
9 Creative Commons. CC0. Geraadpleegd op 16-2-2022.

2. Slimme algoritmiek om krantenfoto's en persfoto's te matchen

In stappen naar een model voor beeldherkenning

Het herkennen van beelden werkt bij computers heel anders dan bij mensen. In feite gebruiken mensen de helft van de kracht van onze hersenen om patronen, vormen en texturen om ons heen te herkennen.

Computergestuurde patroon- en objectherkenning begint echter met een reeks geordende pixels (meer dan een miljoen), met kleurenwaarden van nul (donker) tot 255 (helder). Kranten, gedrukt volgens het halftoonprincipe, maken gebruik van verschillende maten zwarte inktstippen om de grijsgradaties weer te geven.

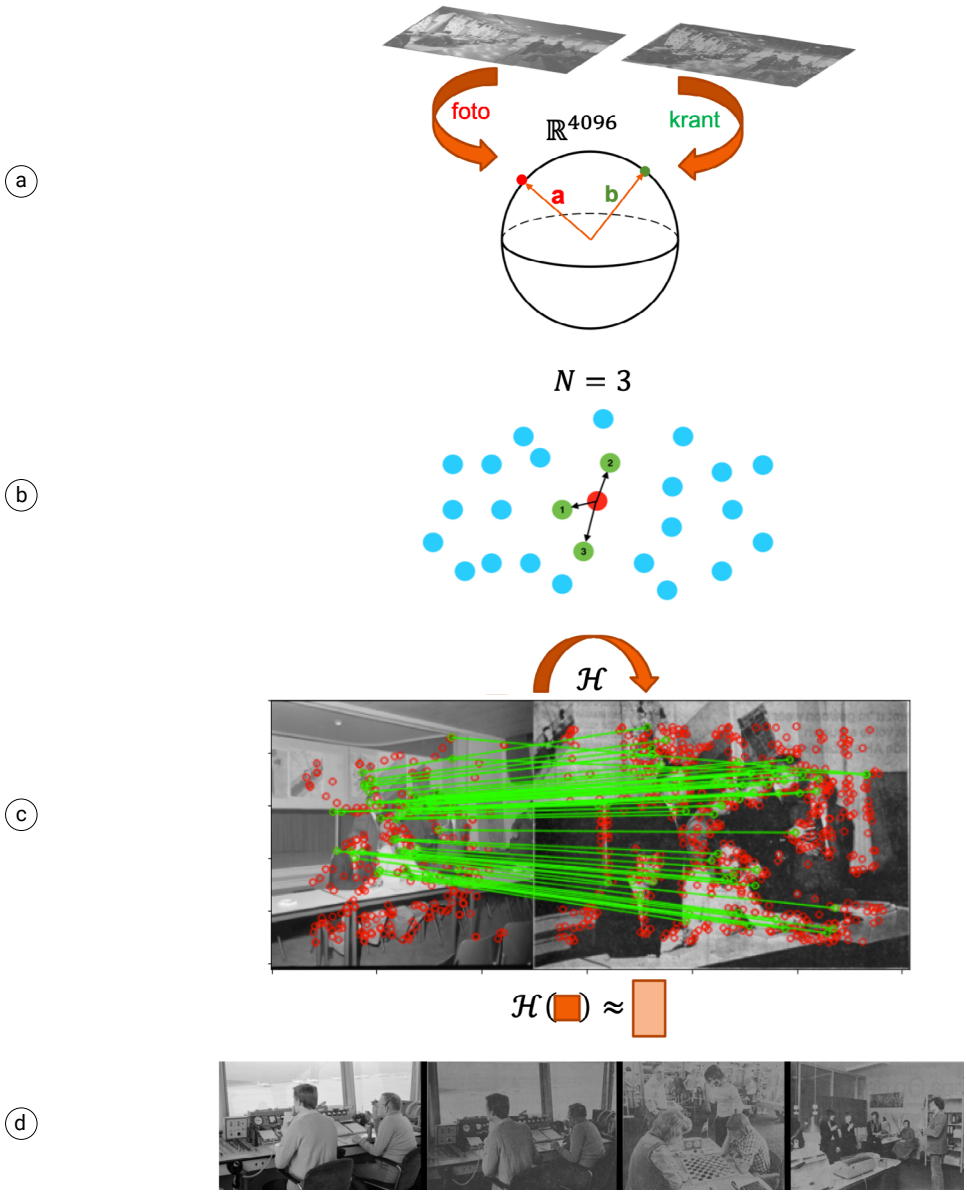
Links een foto in de persfotocollectie, rechts dezelfde in de krant. De uitdaging was om deze via beeldherkenning te verbinden. Gasunie schaaktoernooi in Groningen (1975). Foto: Persfotobureau D. van der Veen. CC0.
Gepubliceerd in *Nieuwsblad van het Noorden*, 22 december 1975, p. 31. Collectie: Groninger Archieven.

Om digitale foto's uit persarchieven met krantenillustraties te vergelijken, is een algoritme ontwikkeld dat de informatie op een laag niveau integreert en vervolgens gebruikmaakt van deze representatie van de afbeelding op hoog niveau, vergelijkbaar met wat mensen doen. Een logische kandidaat voor het oplossen van dergelijke problemen is een diep convolutienetwerk zoals VGG16[10], getraind op meer dan een miljoen afbeeldingen om duizend verschillende categorieën te classificeren. Het is niet alleen bruikbaar voor classificatie, maar ook om uit meerdere schalen en diepten van het netwerk een vector van 4.096 globale en enigszins abstracte kenmerken te extraheren.

Om krantenillustraties en foto's te kunnen vergelijken, zijn de bronbestanden eerst geblurd door middel van een Gauss-filter, waarna normalisatie en schaling zijn toegepast. Om vervolgens gelijkwaardige foto's te vinden, kan de Euclidische afstand tussen de kenmerkvectoren berekend worden. Op deze

10 Simonyan, K. & Zisserman, A. 'Very Deep Convolutional Networks for Large-Scale Image Recognition' (2014). In principe kan men elk ander convolutienetwerk gebruiken om abstracte kenmerken uit de afbeeldingen te extraheren, bijvoorbeeld VGG19 of ResNet.

manier worden voor elke persfoto de N-dichtstbijzijnde buren[11] onder de krantenfoto's gevonden. Hoe kleiner de afstand tussen een paar kenmerkvectoren, hoe groter de kans dat twee afbeeldingen vergelijkbaar zijn, tenminste in deze hoogdimensionale kenmerkruimte.



Deze processtappen en technieken zijn met opzet in het Engels ten behoeve van leesbaarheid:
a) Mapping on images to feature space.
b) Finding three nearest neighbours in feature space with fast search computing Euclidean distances.
c) Matching scale-invariant features between two images. The homography H between two images is computed after excluding outliers with RANSAC.
d) The photograph (left) followed by three best matches from the newspaper, starting with a 'true' match (when rectangle transforms into itself under H).
Afbeelding: Sioux Technologies.
Foto's: Persfotobureau D. van der Veen / Fotopersbureau De Boer. Collecties: Groninger Archieven / Noord-Hollands Archief.

Met deze generieke aanpak vindt het netwerk zoals verwacht de beste matches, altijd binnen de juiste categorie, zoals auto's, gebouwen, mensen en sportevenementen. Bovendien zijn de voorspellingen robuust voor schaalvergroting, bijsnijden en rotaties van afbeeldingen. Convolutienetwerken zijn hierin superieur aan andere op afbeeldingen gebaseerde technieken.

11 We hebben de scikit-learn python-bibliotheek gebruikt om de boom met kenmerkvectoren voor krantenillustratie te construeren en voor elke foto de N-dichtstbijzijnde buren (de N-beste matches binnen deze boom) te lokaliseren. De berekening gebeurt in een fractie van een seconde, in principe onafhankelijk van de waarde van N.

Toch kunnen de geëxtraheerde globale kenmerken en berekende afstanden in de kenmerkruimte geen antwoord geven op de belangrijke vraag of er onder de gevonden beste matches een 'echte' match is tussen de foto en de bijbehorende krantenillustratie – met andere woorden: of het vastgestelde gekoppelde paar (beste match) daadwerkelijk naar dezelfde gebeurtenis wijst. Om deze vraag te beantwoorden wordt een nabewerking uitgevoerd, bestaand uit twee stappen:

Stap 1. Zoek een transformatie[12] tussen de foto en elke kandidaat krant (voor de drie beste matches). Om deze transformatiematrix te berekenen is het SIFT-algoritme[13] gebruikt, dat lokale kenmerken in beelden detecteert.

Stap 2. Controleer of bij de bovenstaande transformatie de oriëntatie gehandhaafd blijft en niet vervormd is. Als dit zo is, zijn de foto en krantkandidaat een echte match.

Aan een dergelijke globale geometrische beperking is zeer moeilijk te voldoen, tenzij de afbeeldingen voldoende lokale overeenkomsten hebben. Daarom zijn er bijna geen vals-positieven gevonden door het algoritme en is >99% van de gesuggereerde 'echte' matches daadwerkelijk correct.

De resultaten

In de Groninger Archieven, voor de periode 1970-1979, vond het algoritme 11.853 'echte' matches. Dat betekent dat 3,5% van alle persfoto's een betrouwbare koppeling heeft met een krantenfoto. Voor het Noord-Hollands Archief is het aandeel 'echte' matches met meer dan 18.000 foto's hoger (4,6%). Het aantal gekoppelde fotoreportages ligt hoger, omdat iedere fotoreportage uit één of meerdere beelden bestaat. Hiermee is aangetoond dat een voor mensen omslachtige taak binnen een beperkte tijd geautomatiseerd is uit te voeren.

Collecties NHA	Testdata	Verbonden afbeeldingen	Verbindingen met hoge waarschijnlijkheid
Krantenpagina's	166.776	392.917	
Persfoto's	410.879	405.599	18.770

Collecties GrA	Testdata	Verbonden afbeeldingen	Verbindingen met hoge waarschijnlijkheid
Krantenpagina's	95.620	157.791	
Persfoto's	90.597[14]	337.453	11.853

De gevonden matches zijn in een CSV-bestand opgenomen en zijn nodig voor de presentatie in de demo. Per foto (foto-ID) zijn de volgende gegevens vermeld:

- Match met de desbetreffende krant (ID)
- Link naar de krant
- Link naar de thumbnail voor de krant
- Coördinaten van de plaats van de foto op de krantenpagina

Juridische aspecten bij het toepassen van beeldherkenning

Beeldherkenning toepassen en laten toepassen op foto's van herkenbaar afgebeelde, levende mensen is een verwerking van persoonsgegevens. De risico's daarvan moeten in kaart worden gebracht in een Data Protection Impact Assessment (DPIA). Betrokkenen moeten worden door de verwerkingsverantwoordelijke worden geïnformeerd in een privacyverklaring. Elke instelling is hierbij zelfstandig verantwoordelijk voor de data uit de eigen collectie.

12 Een transformatie is een wiskundige term die vastlegt hoe je van punt(en) in situatie 1 naar punt(en) in situatie 2 komt. Dit kan o.a. in tijd en ruimte zijn. https://nl.wikipedia.org/wiki/Transformatie_(wiskunde)

13 Scale-invariant feature transform; https://nl.wikipedia.org/wiki/Scale-invariant_feature_transform.

14 Het gaat hier om negatiefstroken, met 2-4 foto's per strook.

3. Het zoeken door de foto's: de demo en de gebruiker

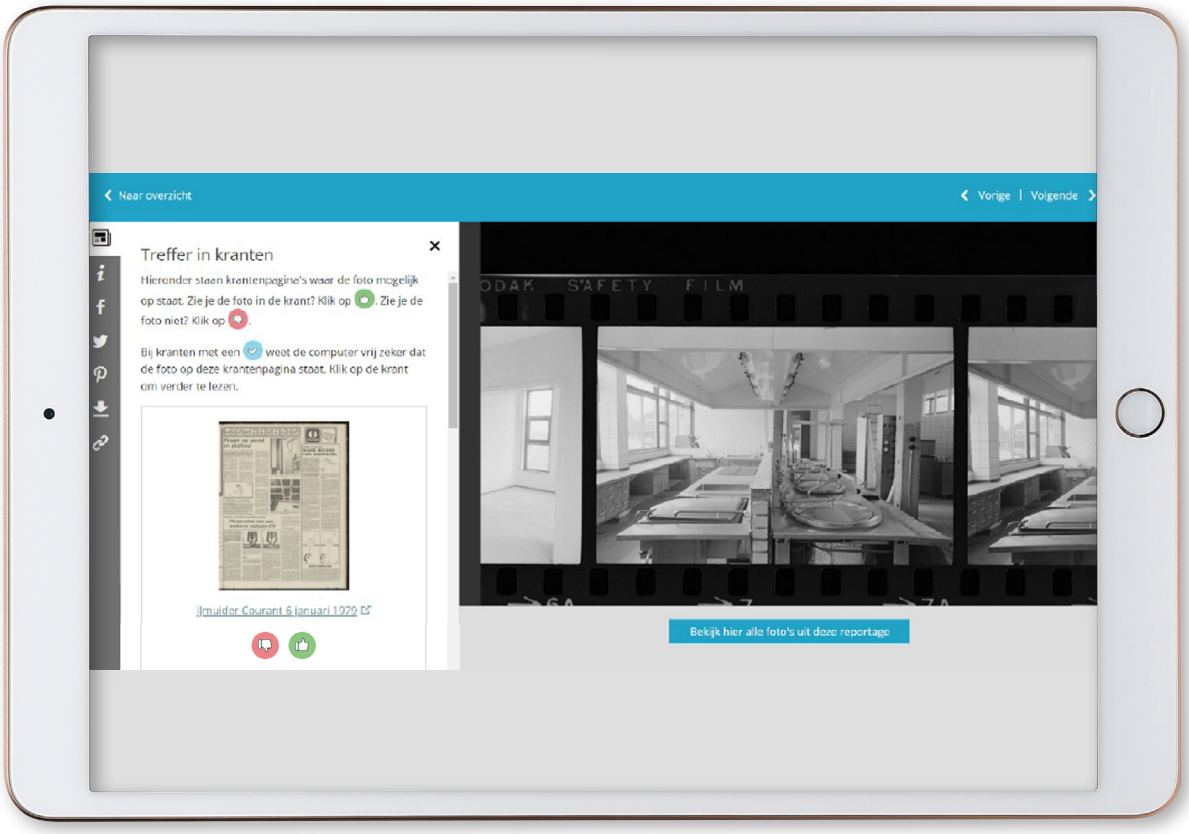
De demonstrator: een interface als proof of concept

In hoofdstuk 2 beschreven we dat in een CSV-bestand de gemaakte koppelingen tussen persfoto's en krantenfoto's zijn vastgelegd. Deze data zijn opgeslagen bij de aanwezige metadata van het foto- of reportagerecord. De overgrote meerderheid van de fotorecords had drie mogelijke matches; daarom diende de uitbreiding van het record herhaalbaar te zijn. In het geval van een complete fotoreportage waren er drie matches voor iedere afzonderlijke afbeelding op de reportage. Dit resulteerde in een veelvoud van drie. Ter illustratie: bij tien foto's in een reportage was het aantal mogelijke matches dertig.

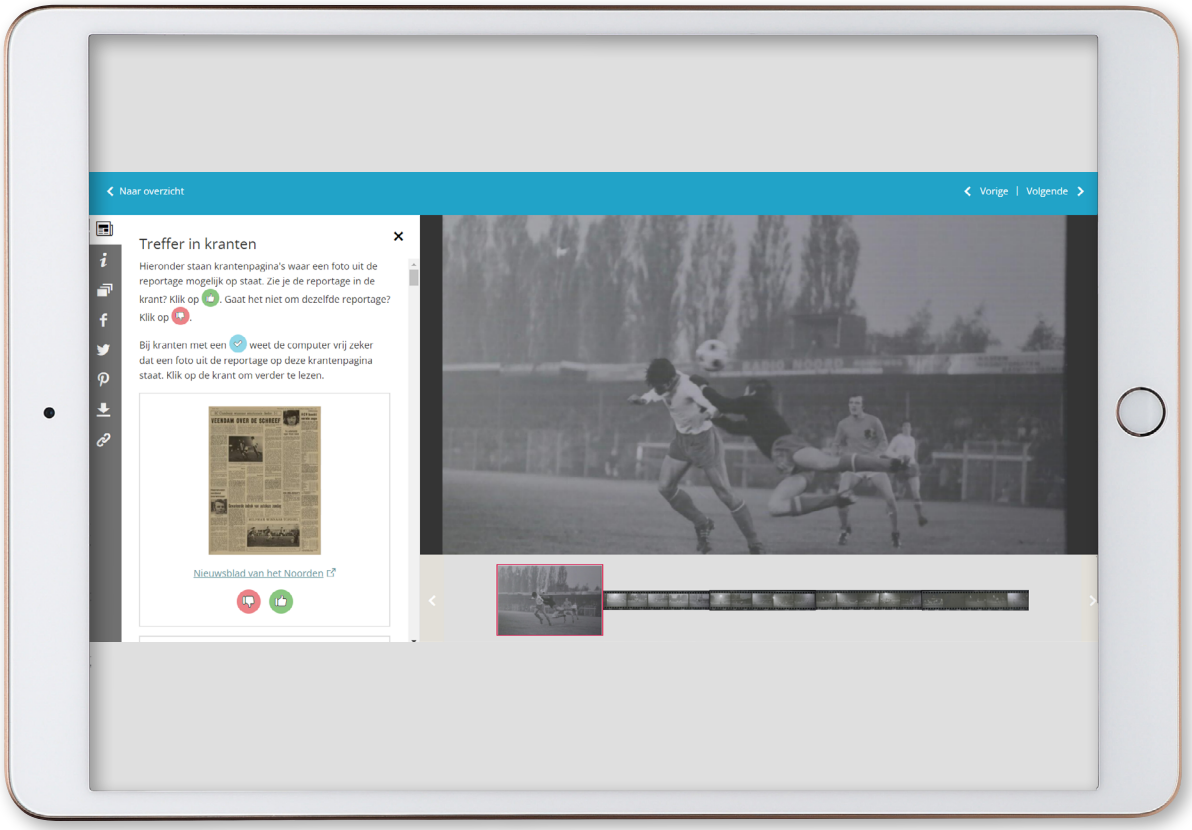
Gezien het doel van het project moest in een oogopslag duidelijk zijn dat een foto een aantal mogelijke matches kon hebben met een afdruk in een krant. Gekozen is om de kranten naast de foto te tonen in de vorm van thumbnails en een link, in plaats van een keuze tussen deze twee te maken. Het toevoegen van thumbnails onder de foto tonen levert problemen op met scrollen, voornamelijk bij kleinere beeldschermen. Ook door het duidelijk scheiden van de metadata en de thumbnails van de matches kwam de nadruk meer op 'krant en foto's' te liggen. Daarnaast maken de thumbnails het beeldrecord visueel aantrekkelijk. Voor het scheiden van de metadata van de kranten is gekozen voor verschillende tabbladen, waarbij het default tabblad geopend staat op de kranten.

De reportage achter de persfoto

Een ander wezenlijk onderdeel van het project is het kunnen bekijken van de complete reportage. Deze is onder de afbeelding getoond als opvallende link, of, wanneer mogelijk, als scrolbalk.



Een reportage getoond als link onder de foto. Met één muisklik beland je in de complete reportage.
Screenshot krant-en-fotos.nl.



Een reportage, bestaande uit meerdere beelden, in zijn geheel weergegeven.
Screenshot krant-en-fotos.nl.

Onder de thumbnail staat de link van de krant weergegeven. De krantnaam en de publicatiedatum worden uit de URL gefilterd. In het geval van Delpher komt de info uit de mpeg21-didl, het bestandsformaat waarin de metadata worden beschreven.

Matches door de computer én de gebruiker

Zoals in hoofdstuk 2 beschreven, vindt het algoritme in zo'n 4% van de gevallen een 'echte' match. Het gaat om foto's waarbij de computer meer dan 99% zeker weet dat deze overeenkomt met een foto uit een krantenpagina. De kans dat de gebruiker een 'echte' match vindt, is dus zo'n 4%. Omdat het doel van de beeldbank is om de gebruikerservaring van aan persfoto's gekoppelde krantenverhalen te peilen, staat de beeldbank standaard gefilterd op 'grote kans op treffer'. Een grote kans op een treffer wordt bij een persfoto aangegeven met een vinkje. Voor het overige deel bieden we bezoekers de mogelijkheid om de gevonden match te beoordelen aan de hand van duimpje omhoog of omlaag. De score daarvan wordt bewaard in de backend van de website. De werking hiervan is een simpele teller. Elke keuze wordt bevestigd als +1. Om ervoor te zorgen dat een bezoeker slechts eenmaal per foto kan aangeven of het wel of geen match betreft, is een aantal opties overwogen:

- **Cookies instellen**. Aangezien het gaat om cookies voor de werking van de site, is het geen probleem om deze te zetten. Mogelijk blokkeren bezoekers alle cookies in hun browser. In dat geval kunnen gebruikers meerdere keren achter elkaar stemmen.[15]
- **Een sessiecookie.** Sessiecookies blijven standaard 24 minuten na de laatste aanroep bestaan. Ze blijven alleen actief zolang de browser openblijft en zolang de gebruiker actief is op de website. Daarna verdwijnen de gegevens automatisch weer. Sessiecookies zullen ook minder snel worden tegengehouden door een cookieblocker.

15 https://veiliginternetten.nl/thema/privacy/cookies-browsegeschiedenis-en-auto-aanvullen/moet-ik-een-cookiemelding-op-mijn-website-zetten/

- **Een cookie op het IP-adres**. Deze oplossing is de minst ideale, omdat meerdere gebruikers op hetzelfde IP-adres kunnen werken, bijvoorbeeld op een kantoor. Het zou dan nodig zijn om de stemming per stuk en per IP-adres op te slaan. Vanwege privacywetgeving is het niet wenselijk deze data op te slaan.[16]

Gekozen is voor de sessiecookie. Hieraan kleven de minste risico's als het gaat om het blokkeren van cookies. Ook worden geen andere gegevens opgeslagen dan gekozen zijn.

Een gebruiker kan hier een treffer aangeven door op het groene duimpje te klikken.
Screenshot krant-en-fotos.nl.

Keuzes in de presentatie in de demo

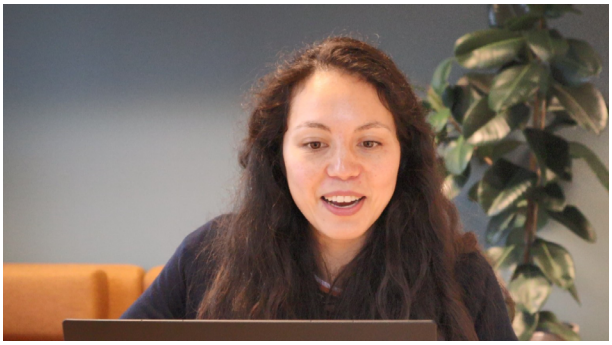
Een uitdaging bij het maken van de demonstator was het samenbrengen van meerdere collecties met verschillende structuren, afkomstig vanuit diverse erfgoedinstellingen. Zo is een keuze gemaakt in het tonen van doorzoekbare metadata van de persfotocollecties. Alleen filteropties zijn gebruikt die voor beide collecties relevant (gevuld) zijn. Ook de manier van presentatie van de persfoto's verschilt. De foto's van Fotopersbureau De Boer worden getoond per unieke foto, met voor iedere foto een kranten-koppeling. Die van Persfotobureau Van der Veen worden gebundeld per fotoreportage getoond, met daarbij ook gebundeld alle mogelijke krantenkoppelingen van die reportage (zie de schermafdrukken hierboven).

Omdat dit gevolgen heeft voor de demonstrator, is het aan te bevelen om in een vroeg stadium na te gaan waar samen te brengen collecties raakvlakken hebben en waar ze verschillen. Ook tijdens en vlak na de bouw van de demonstator zijn nog aanzienlijke aanpassingen gedaan. Bij het starten van een vergelijkbaar project is het aan te bevelen om hiervoor voldoende tijd in te ruimen.

Gebruiksonderzoek

Om te weten te komen hoe de demonstrator wordt ervaren door eindgebruikers, is een gebruiksonder-zoek uitgevoerd. Hierbij is meegekeken bij de schermervaring van de demonstrator volgens de *think aloud*-methode.[17] Acht personen deden mee, waarvan vijf personen uit de omgeving van Haarlem en drie uit de omgeving van Groningen. Hoewel zij een diverse achtergrond hadden – jong en oud, man en vrouw, weinig en veel ervaring – waren alle deelnemers potentiële gebruikers. Zo deden een biblio-thecaris, wetenschapper, oud-persfotograaf, vrijwilliger en beheerder van een Facebookpagina waarop oude foto's centraal staan mee. Zeven gebruikers gaven aan enige ervaring te hebben met digitaal zoeken in historische fotocollecties, zoals de online beeldbank van de archiefinstelling. Zes gebruikers hadden enige ervaring met het digitaal zoeken in historische kranten, bijvoorbeeld via Delpher.

16 Internetrecht door Arnoud Engelfriet. Mag IP-adressen loggen van de privacywet? Geraadpleegd op 16-2-2022.
17 De hardopdenk-methode.



Een van de deelnemers tijdens het gebruiksonderzoek (2022).
Foto: Noord-Hollands Archief.

Het gebruiksonderzoek leverde vijf leerpunten op:

1. **Iedereen zoekt anders**. Waar de ene gebruiker in de demonstrator vooral zoektermen invoert, gaat een ander aan de slag met de filtermogelijkheden en hanteert een derde allerlei varianten daartussen. De gebruikte zoekmethode heeft invloed op de kans op een positief resultaat, net als de kwantiteit en kwaliteit van de beschikbare doorzoekbare metadata. Een (visuele) zoekhulp kan een gebruiker assisteren, maar uit het onderzoek blijkt ook dat ongeveer de helft van de gebruikers de nu geboden uitleg bij de krantenmatches niet leest, of slechts nadat erop is gewezen.
2. **'Herrieschoppers' niet gewenst.** Gebruikers willen alleen krantenkoppelingen zien met een 'grote kans op een treffer'. In dit project is gefocust op treffers met meer dan 99% kans, maar een andere hoge grenswaarde zou ook mogelijk zijn. Andere geboden krantentreffers, met een lage kans, worden door gebruikers niet aandachtig bekeken of als verwarrend ervaren. Eén gebruiker wees op de wenselijkheid van 'een beperking in een woud van informatie', een ander noemde de treffers met een lage kans 'herrieschoppers'. De mogelijkheid om aan te kunnen geven dat een treffer onjuist is via een 'duimpje omlaag', ook bij een 'grote kans op een treffer', wordt als positief en laagdrempelig ervaren. Wel is het de vraag of gebruikers daadwerkelijk op de duimpjes zullen klikken. 'Ik kan me voorstellen dat mensen zo gefocust zijn op het vinden van datgene waar ze naar zoeken, dat ze die duimpjes vergeten en dat ze denken: ik blijf niet bezig,' aldus een gebruiker.
3. **Weergave nog niet goed genoeg.** De foto's van Fotopersbureau De Boer worden getoond per unieke foto, met voor iedere foto een krantenkoppeling. Die van Persfotobureau Van der Veen worden gebundeld per fotoreportage getoond, met daarbij ook gebundeld alle mogelijke krantenkoppelingen van die reportage. Beide weergaven worden door gebruikers niet als optimaal ervaren.
    - Als beelden worden getoond per unieke foto, hebben gebruikers moeite om alle foto's uit de reportage terug te vinden. De knop 'Bekijk hier alle foto's uit deze reportage' wordt door vier van de acht gebruikers over het hoofd gezien.
    - Als beelden gebundeld per reportage worden getoond, hebben gebruikers moeite om de gebundelde krantenkoppelingen te interpreteren. In een door ons voorgelegd voorbeeld kwam de eerst getoonde persfoto uit de reportage niet overeen met de gekoppelde persfoto in de krant; het betrof een andere foto uit de gehele reportage. Ten onrechte klikten vier van de acht gebruikers op het 'duimpje omlaag'.
4. **Duidelijkheid koppelingen.** Voor gebruikers is het onduidelijk hoe de door het algoritme gemaakte koppelingen tot stand zijn gekomen. Zo lijkt een aantal van de gebruikers te denken dat de koppelingen niet tot stand zijn gekomen door beeldherkenning, maar door tekstuele overeenkomsten tussen de beschreven fotoreportages en de teksten in de krant. Het plaatsen van een kader om zowel de gematchte persfoto als de vondst van diezelfde foto in de krant kan een oplossing zijn.
5. **Positief over functionaliteit.** Alle gebruikers zijn positief over de getoonde resultaten van de demonstrator en de achterliggende techniek. Allen geven aan positief te staan tegenover integratie van de functionaliteit in de bestaande beeldbanken van archiefinstellingen. Dat de computer ook fouten maakt in de koppelingen wordt als aanvaardbaar ervaren, zolang wordt aangegeven dat de koppelingen door een computer op basis van waarschijnlijkheid worden gemaakt. Een gebruiker

zegt bijvoorbeeld over het geheel: 'Ik ben onder de indruk. Ik vind het mooi, het ziet er leuk uit, het werkt goed. Ik denk dat jullie een hoop mensen hier blij mee maken. Die koppeling met de kranten is gewoon superhandig. En dat het niet 100% sluitend is, is prima; misschien stoot je een keer je neus. Maar al die andere keren dat je wat vindt, op zo'n eenvoudige manier, dat is prettig.'

Juridische aspecten bij het ontwikkelen van een interface

Ook in de ontwikkeling van de interface moesten we rekening houden met de beperkingen op hergebruik zoals gesteld in de contracten tussen de erfgoedinstellingen en CBO's en uitgevers. Op de website *Krant en Foto's* worden de foto's uit de persfotoarchieven getoond samen met een thumbnail van de gehele krantenpagina met daarop de mogelijke match. Die krantenpagina bevat naast foto's ook artikelen die auteursrechtelijk beschermd zijn. De thumbnails dienen daarom dermate klein te zijn dat de tekst niet leesbaar is, zodat het niet als nieuwe beschikbaarstelling beschouwd kan worden, waarvoor opnieuw toestemming nodig zou zijn van de CBO's en uitgevers.

Daarnaast sluiten de instellingen een overeenkomst af tussen verwerkingsverantwoordelijken. Daarin staat onder andere dat er een privacyverklaring moet komen op de website krant-en-fotos.nl, die verwijzingen zal bevatten naar de privacyverklaringen van de individuele instellingen.

4. Resultaten en inzichten: een samenvatting

De resultaten

Het project *Krant en Foto's* heeft diverse resultaten opgeleverd, die te vinden zijn via de website krant-en-fotos.nl:

- Een werkend model voor het verbinden van foto's op basis van beeldherkenning. De software voor de beeldherkenning is open source beschikbaar gesteld onder Apache 2.0-licentie via https://github.com/siouxmathware/match_newspaper_photos.
- Een demonstrator: een beeldbank met een collectie scans van krantenpagina's en persfoto's, waardoorheen je kunt zoeken. De persfoto's die je vindt zijn verbonden aan krantenfoto's en aan fotoreportages.
- Een analyse van ervaringen van gebruikers met de demonstrator. Naast een samenvatting in dit whitepaper zijn de resultaten terug te zien in een korte video.

Inzichten

Dit project heeft bewezen dat het mogelijk is om met kunstmatige intelligentie in een grote dataset met de gebruikte technieken foto's aan elkaar te koppelen. Het laat zien wat AI kan betekenen voor een productieproces binnen de cultuur- en mediasector. Daarom is het ook opgenomen als use case van de Werkgroep Cultuur en Media van de Nederlandse AI Coalitie.[18]

We hopen dat de lessen en resultaten zullen landen in de erfgoedsector en dat professionals een duik zullen nemen in de beeldbank. Gericht zoeken of juist rondstruinen kan leiden tot onverwachte ontdekkingen en nieuwe verhalen op basis van de artikelen en foto's in de kranten aan de ene kant en de fotoreportages aan de andere.[19]

In het geval dat andere erfgoedinstellingen zelf een project gaan opzetten waarin kranten, foto's of beeldherkenning een rol spelen, kunnen de volgende lessen van pas komen.

Dataverzameling

- Zelfs als dataverzamelingen op elkaar lijken is het verstandig te kijken naar je procesdoelen en -stappen. Kijk vervolgens welke vorm en inhoud voor je data en metadata noodzakelijk zijn en of conversies nodig zijn. Nog beter is het als de erfgoedsector samen met netwerkpartners verder werkt aan standaardisatie. In dit project leek de opbouw van metadata van de beide persfotocollecties sterk op elkaar, maar in de data zat een groot verschil: bij Van der Veen bevatte elke scan een negatiefstrook met meerdere beelden, terwijl het bij De Boer om losgeknipte beelden ging.
- Juridische aspecten zijn van grote invloed op de mogelijkheden om data uit te wisselen tussen partijen, te verwerken en te publiceren. Als verschillende partners samenwerken is het belangrijk te kijken naar eigenaarschap van de data en gemaakte afspraken over de data. Voor het verwerken van persoonsgegevens van partner 1 door partner 2 is een verwerkersovereenkomst nodig. Van belang is te kijken wat mag met de data volgens de wet, afspraken met CBO's, uitgevers en andere stakeholders.

18 Nederlandse AI Coalitie. Toepassingsgebieden. Geraadpleegd op 16-2-2022.
19 Noord-Hollands Archief. Bij deze foto begint mijn jeugd: verhaal achter de foto. Geraadpleegd op 16-2-2022.

- Het oorspronkelijke idee binnen dit project was om het neurale netwerk te trainen op door mensen gemaakte paren van foto's en kranten. In de menselijke interpretatie van beeldovereenkomsten bleek echter dat aanzienlijke ambiguïteit te zitten. Het uiterlijk van Sinterklaas die per boot arriveert, wordt bijvoorbeeld geassocieerd met een andere gebeurtenis dan Sinterklaas op zijn paard. Dit is echter context-gebaseerd en geen daadwerkelijke afbeeldingsovereenkomst.
- Om een eenduidige *ground truth*-dataset te creëren, moeten het concept en de maatstaf van wat een beeldovereenkomst is, nauwkeurig worden gedefinieerd. Moet de in de krant gepubliceerde foto *exact* hetzelfde zijn als degene uit de persfotocollectie? Of is een foto van dezelfde gebeurtenis, maar net uit een andere hoek, ook voldoende?

Algoritme

- Aan de kant van het algoritme hebben de geleerde lessen onder andere betrekking op de eerste stap van het voorbewerken en extraheren van individuele illustraties uit de krant met behulp van de OCR-tool van ABBYY.[20] Dit proces moet worden verfijnd, net als het algoritme voor het extraheren van foto's uit de negatiefstroken, om een zuivere invoer voor het algoritme te creëren. Maak daarbij maximaal gebruik van metadata, zoals jaren en uitgever, om een zoekruimte van mogelijke overeenkomsten te beperken en een snellere algoritmeprestatie mogelijk te maken.
- Het vooruitplannen van de rekenmiddelen, infrastructuuranalyse voor grote datasets en parallel computergebruik zal ook helpen om binnen een acceptabele tijd resultaten te krijgen. Anders duurt het circa 4 seconden per foto om de N=3 beste matches te verkrijgen.
- Om meer matches te vinden tussen de kranten- en fotocollecties, kan het algoritme ten slotte verder worden uitbreid naar andere, meer specifieke kenmerken, zoals gezichtsherkenning, of juist door de flexibiliteit van context-matches toe te laten.

Demo en gebruiksonderzoek

- Voor de beschikbaarstelling van auteursrechtelijk beschermd materiaal moeten afspraken worden gemaakt met rechthebbenden. Het is belangrijk dit van tevoren te regelen, zowel voor de interface zelf als voor de krantenwebsites waarnaar verwezen wordt. Voor de laatste moet dit gebeuren zodat niet naar ontoegankelijke websites wordt gelinkt met materiaal dat alleen op locatie van de erfgoedinstelling te raadplegen is. Voor de interface moet dit gebeuren omdat de auteursrechtelijk beschermde foto's en de rest van de krantenpagina in een grote thumbnail worden getoond. Thumbnails zijn in zeer klein formaat te tonen zonder afspraken. Wanneer deze groter worden, kan dit als afzonderlijke beschikbaarstelling van het auteursrechtelijk beschermde werk worden gezien. Daarvoor is toestemming van rechthebbenden nodig.
- Een interface bouwen verdient net zoveel aandacht als de nieuwe technieken om erfgoedmateriaal te ontsluiten. Het is belangrijk gebruikers eenduidige informatie voor te schotelen en zaken weg te laten of naar de achtergrond te verdringen die niet bijdragen aan de doelen van de dienst of gebruikers. Een doordacht gebruiksonderzoek kan helpen om daarvoor de juiste ingrediënten te verzamelen. Het vraagt tijd en capaciteit om de resultaten te verwerken in de interface.

20 ABBYY FineReader is commerciële software voor optische tekenherkenning (OCR).

Epiloog
Uitdagingen voor de toekomst

Er zijn veel meer persfotocollecties dan De Boer en Van der Veen. Wij willen erfgoedinstellingen inspireren om actief aan de slag te gaan met kunstmatige intelligentie om hun eigen digitale collecties op nieuwe manieren te ontsluiten en te verbinden, waardoor vervolgens meer mensen die prachtige schatten zullen gaan bewonderen. Het is het onderzoeken waard of het daarbij mogelijk is om ook andersoortige foto's, illustraties of gedigitaliseerde objecten uit beeldbanken met elkaar te verbinden.

Verbeteringen van de resultaten binnen Krant en Foto's

De partners binnen het project kijken met tevredenheid terug op de behaalde resultaten binnen een kortlopend traject. Tegelijk zien we dat de producten die we ontwikkeld hebben nog kunnen worden verbeterd. Die verbeteringen kunnen niet meer binnen dit project gerealiseerd worden. Toch delen we ze graag, met het oog op het voornaamste doel van het project: het breder verspreiden van kennis.

Linken vanuit krantendatabases naar de persfoto's

De aangeleverde coördinaten van de foto op de krantenpagina zijn niet gebruikt voor de demonstrator. Deze zouden in de toekomst gebruikt kunnen worden om de foto op de krantenpagina te omkaderen. Op deze manier is het niet alleen mogelijk om vanuit de beeldbank direct naar de juiste foto op een krantenpagina te gaan, maar wordt het ook mogelijk om vanuit Delpher of de NHA-krantenviewer van de foto in de krant naar de complete reportage op een beeldbank te gaan. Hiervoor is bij de ontwerpfase van het project een mock-up gemaakt. Met een kader op de scan van de krant wordt het voor een gebruiker van de demo ook duidelijk welke foto op een krantenpagina zou moeten matchen met de persfoto.

Van de krant in krantendatabases doorklikken naar de foto's in de persfotocollecties is een mogelijke verbetering. *Haarlems Dagblad*, 26 maart 1974, p. 1. *Screenshot mock-up NHA-Krantenviewer.*

Resultaten uit gebruiksonderzoek verwerken in de demonstrator
Een groep gebruikers heeft gekeken of de demonstrator begrijpelijk is opgebouwd en of men aan de hand van opdrachten kon uitkomen op de juiste locatie in de beeldbank. Deze resultaten geven aanleiding tot allerlei aanpassingen aan de demonstrator. Ze zouden kunnen zorgen voor een betere gebruikerservaring.

Matches van gebruikers verwerken in de dataset en demonstrator
Gebruikers kunnen in de beeldbank aangeven of verbindingen die door het model worden gepresenteerd echte matches zijn: gaat het echt om dezelfde kranten- en persfoto? Na klikken op duimpjes worden gegevens opgeslagen. De opgeslagen gegevens zouden kunnen terugvloeien in de demonstrator. Een gebruiker krijgt dan meer juiste verbindingen tussen persfoto's en krantenfoto's te zien.

Hiervoor zal een proces moeten worden ontwikkeld. Wanneer vertrouw je er bijvoorbeeld op dat twee foto's overeenkomen? Hoe geef je in de demonstrator weer dat deze match door gebruikers is vastgesteld en onderscheid je ze van door het model gesuggereerde matches? In hoeverre kun je de nieuwe matches inzetten als extra trainingsdata om het algoritme verder te ontwikkelen?

Frisse ideeën voor de toekomst

Enthousiast geworden professionals in archieven, musea en bibliotheken zouden net als bij *Krant en Foto's* de handen ineen kunnen slaan met partners met expertise op het gebied van technologie en digitaal erfgoed, om zo nog meer persfotocollecties te verbinden aan krantencollecties. Opschalen kan met de collecties De Boer en Van der Veen, maar dan voor een andere periode, of met hele andere fotocollecties. Ook voor kranten liggen nog vele digitale collecties te wachten om op deze manier beter ontsloten te worden.

Initiatieven van NDE en CLARIAH op gebied van standaardisatie kunnen als voorbeeld dienen om data op een zo gestandaardiseerd mogelijk wijze aan te leveren.

Extra aandacht zal bij het opschalen moeten gaan naar het verwerken van de data bij beeldherkenning. *Cloud based computing*, gebruikmakend van state-of-the-art GPU's (grafische processoren) en een geparallelliseerd algoritme, kan zorgen voor een versnelling, waardoor ook grotere datasets beter te verwerken zijn.

Natuurlijk zijn bij een vervolg allerlei afslagen mogelijk. Te denken valt aan het inzetten van kleurenfoto's en te gaan voor recentere kranten en persfoto's. Kleur als aspect meenemen zal waarschijnlijk helpen om te komen tot een hoger percentage verbindingen.

Een andere afslag is het betrekken van de tekst van het krantenartikel bij de beeldherkenning. Uit die tekst is veel metadata te halen, bijvoorbeeld de identiteit van de personen op de foto's, het moment van de opname en de locatie. Die extra metadata kan worden ingezet om de gesuggereerde verbinding tussen twee beelden te verifiëren. Als die extra gegevens worden toegevoegd aan de bestaande beschrijvende metadata of aan nieuw te beschrijven materiaal, kunnen erfgoedorganisaties de metadata via Linked Open Data aanbieden om gebruikers te bedienen en datacollecties verder aan elkaar te koppelen.[21]

Wat de toekomst ook mag brengen voor het geautomatiseerd verbinden van beelden, het is verstandig om bij het opstarten van projecten rekening te houden met alle lessen die in dit paper genoemd worden.

21 Binnen CLARIAH wordt gewerkt aan een generiekere aanpak om verrijkingen herbruikbaar te maken. Het is verstandig hierop aan te sluiten.

Colofon

Dit whitepaper is tot stand gekomen dankzij bijdragen van de partners in het project *Krant en Foto's* van juli 2021 tot maart 2022.

Auteurs (tussen haakjes Orcid)**:** Evgeniya Balmashnova; Rob Bertrams; Martine Brons (0000-0001-7581-2327); Jeroen Franken; Mark Groothuis (0000-0001-5905-0553); Roosmarijn de Groot (0000-0002-5127-4162); Michel de Gruijter (0000-0002-8715-0129); Bram Kampen (0000-0001-5100-5950); Martijn Kleppe (0000-0001-7697-5726); Jan Kruidhof (0000-0002-3798-2488); Oksana Manyuhina (0000-0001-8521-6684); Nico Vriend (0000-0002-6260-6441); Dineke van der Wal (0000-0002-4726-1529)

Eindredactie: Anne van den Dool

Vormgeving: Jeroen Reith (www.burogom.nl)

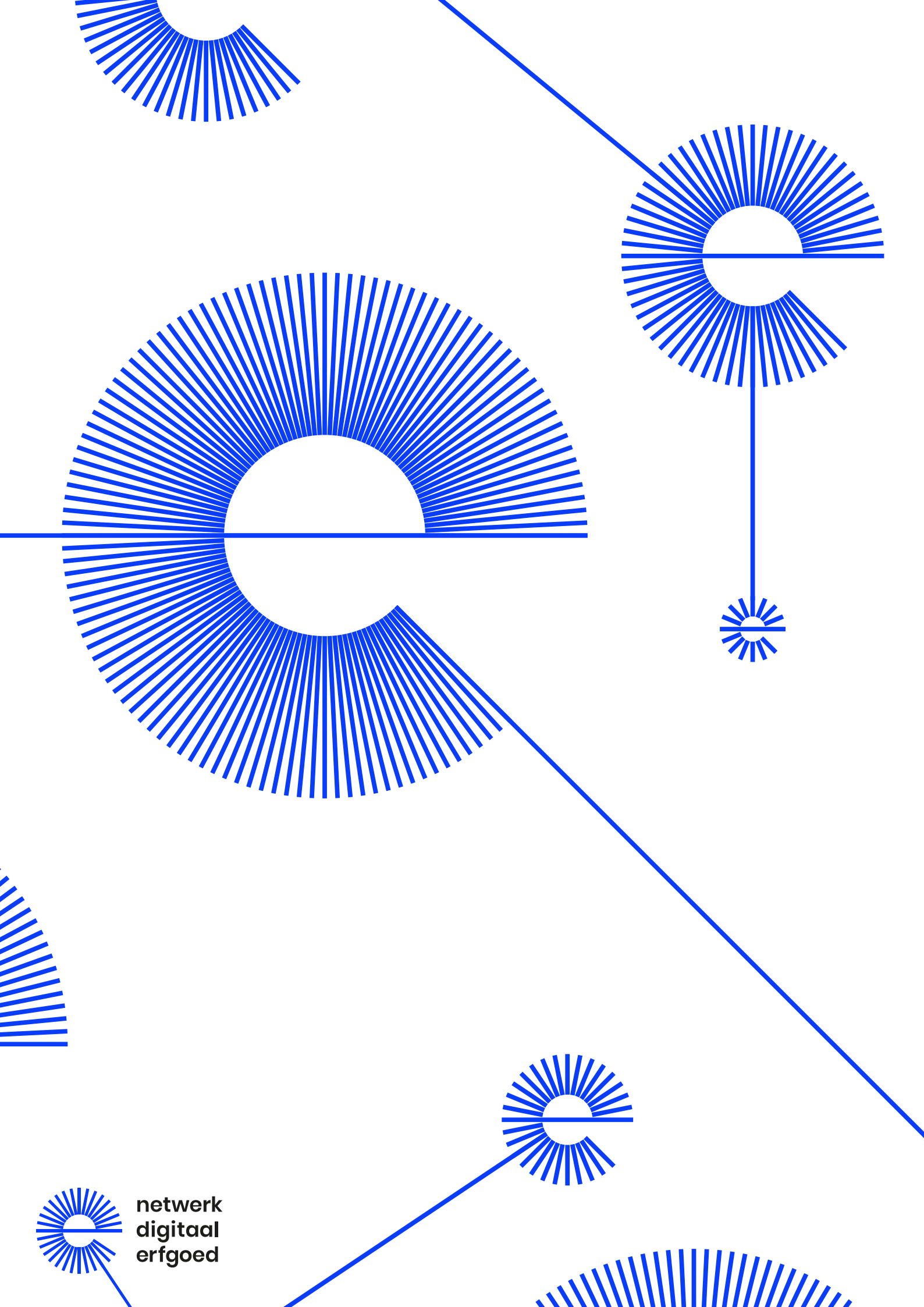
DOI: https://doi.org/10.5281/zenodo.6183002

Foto omslag: Voorbijgangers lezen de krant. Bijkantoor van het Nieuwsblad van het Noorden in de Langestraat te Winschoten, 1972. Foto: Persfotobureau D. van der Veen. Collectie: Groninger Archieven. CC0.

Plaats en jaar van uitgave: Den Haag, 2022

Contact: https://krant-en-fotos.nl / michel.degruijter@kb.nl

Deze publicatie is tot stand gekomen door steun van het ministerie van Onderwijs, Cultuur en Wetenschap en het Netwerk Digitaal Erfgoed.


netwerk
digitaal
erfgoed