#pkdigitaal

Stavaza project: persoonskaarten overledenen 1939 – 1994 digitaliseren

Van karton naar data….

#pkdigitaal

Agenda

1. **Deelresultaat digitalisering**
2. Deelresultaat zoekingang / interne database
3. Afsluitende opmerkingen

CBG

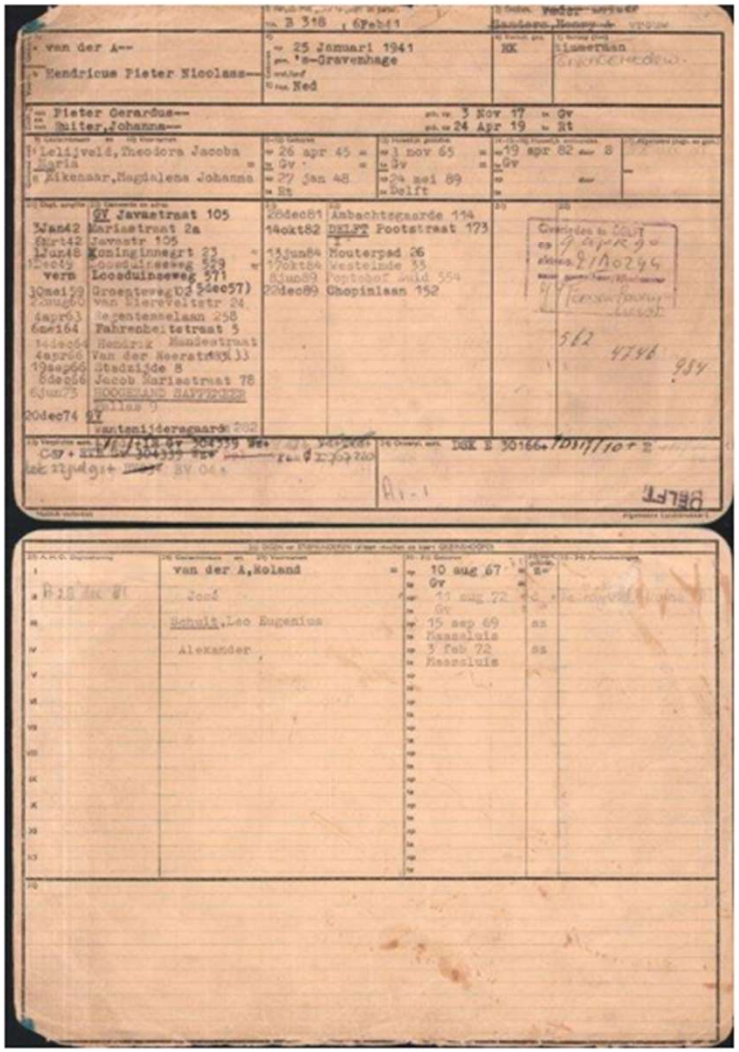
#pkdigitaal

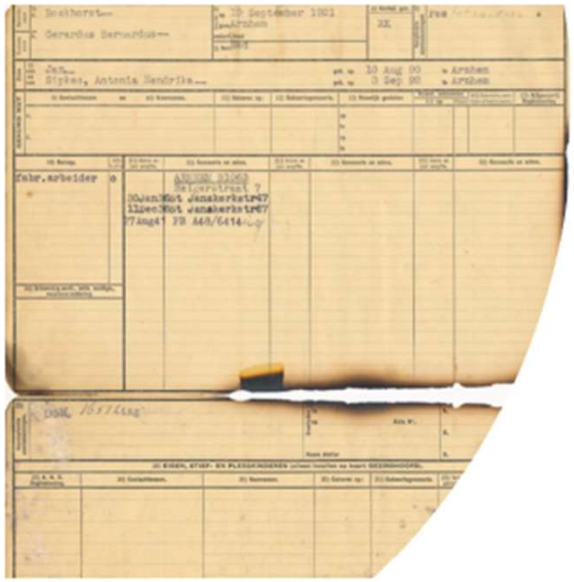
PERIODEN	AANTAL PERSOONSKAARTEN	METER
1939 – 1970	2.880.000	758
1971 – 1980	1.130.000	340
1981 – 1987	920.000	275
1988 – 1994 (tot 1 oktober)	870.000	260
TOTAAL	***5.800.000***	***1.633***

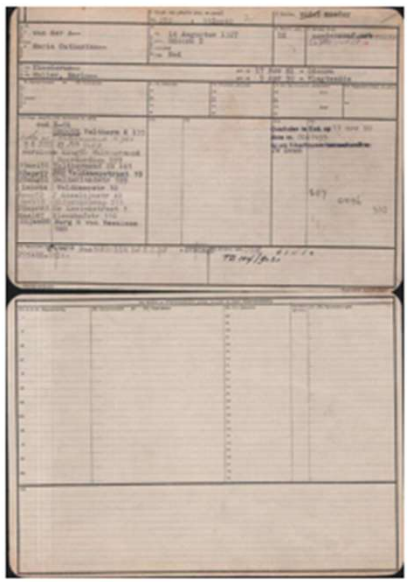
Verhouding standaard persoonskaart / kwetsbare persoonskaart

- Standaard kartonnen persoonskaart 99% = 5.740.000 (met een afwijking van +5/-5%)
- Kwetsbare persoonskaart ruim 1% = 60.000 (met een afwijking van +50/-50%)

CBG







Standaard en kwetsbare persoonskaarten digitaliseren

CBG Centrum voor familiegeschiedenis

#pkdigitaal

Het archief voordat team scanvoorbereidingen aan de slag gaat….

- Iedere persoonskaart wordt gecontroleerd: kwetsbare kaarten worden apart gelegd
- Witte kaarten worden verwijderd
- Plakband, nietjes, etc. worden verwijderd, op vervolgkaart wordt (indien nodig informatie toegevoegd met elek. typemachine)

#pkdigitaal

Het archief nadat team scanvoorbereidingen haar werk heeft gedaan….

- Natuurlijk proberen we de negatieve invloed op de dienstverlening voor de klant zo veel mogelijk te beperken.

#pkdigitaal

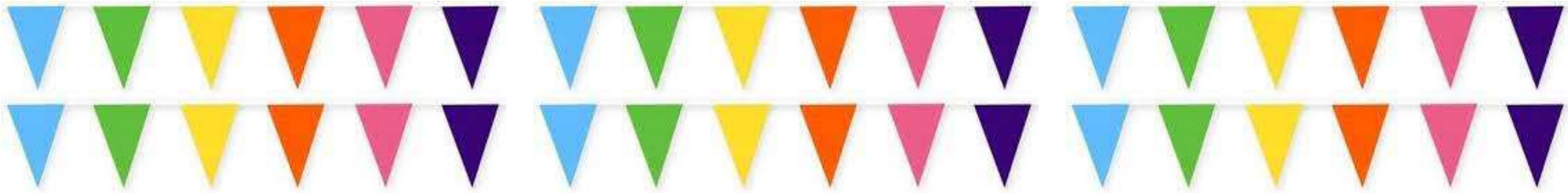
Het archief nadat de dozen weer terugkomen na digitaliseren….

- Iedere doos heeft een etiket met een uniek doos-id en barcode
- De eerste PK van iedere doos wordt gemetadateerd in een google sheet. (autorisatie via Jeroen)
- We gaan vanaf 2/9 naar twee transporten per week (donderdagochtend en donderdagmiddag) wekelijks 200-300 dozen

#pkdigitaal

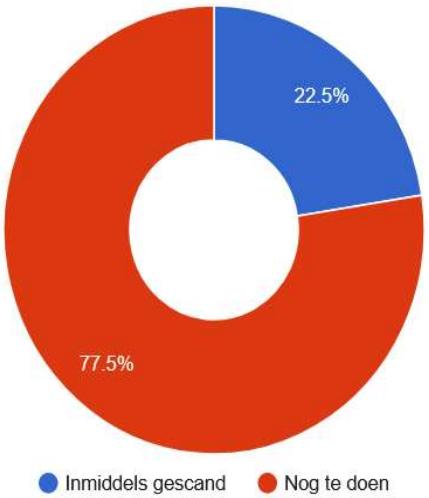
2 september vieren we de mijlpaal:

1 miljoen goedgekeurde persoonskaarten digitaal…



#pkdigitaal

Algehele project status



Totaal aantal kaarten gescand

1.306.936

Dozen

Aantal dozen opgehaald :	**1615**
Aantal dozen retour gebracht :	**1465**
Aantal dozen bij MultiScan :	**150**

#pkdigitaal

Productieschema

26-aug	14	150	830	124.500	1.351.110	
2-sep	15	200	830	166.000	1.517.110	tweemaal
9-sep	16	200	830	166.000	1.683.110	
16-sep	17	250	830	207.500	1.890.610	
23-sep	18	250	830	207.500	2.098.110	
30-sep	19	250	830	207.500	2.305.610	
7-okt	20	250	830	207.500	2.513.110	
14-okt	21	250	830	207.500	2.720.610	
21-okt	22	300	830	249.000	2.969.610	
28-okt	23	300	830	249.000	3.218.610	
4-nov	24	300	830	249.000	3.467.610	
11-nov	25	300	830	249.000	3.716.610	
18-nov	26	300	830	249.000	3.965.610	
25-nov	27	300	830	249.000	4.214.610	
2-dec	28	300	830	249.000	4.463.610	
9-dec	29	300	830	249.000	4.712.610	
16-dec	30	300	830	249.000	4.961.610	
23-dec	31	150	830	124.500	5.086.110	
6-jan	32	200	830	166.000	5.252.110	
13-jan	33	300	830	249.000	5.501.110	
20-jan	34	300	830	249.000	5.750.110	
27-jan	35	120	500	60.000	5.810.110	Kwetsbaar
		7.035		**5.810.110**		

#pkdigitaal

PK beeldbestanden duurzaam bewaren en opnemen in huidige dienstverlening

Masterbestanden duurzaam offline bewaren in een Microsoft Azure datacenter.

Afgeleide bestanden beschikbaar stellen voor huidige dienstverlening.

- Sharepoint omgeving (vervallen)
- Digitaal Depot Multiscan. Het d-depot is een digitale opslagfaciliteit van MultiScan, waarbij op de servers van MultiScan ruimte beschikbaar wordt gesteld aan het CBG. Geautoriseerde medewerkers krijgen via een beveiligde verbinding toegang tot het d-depot door middel van een unieke gebruikersnaam en wachtwoord.
- Censura (door Multiscan ontwikkeld) is een softwareprogramma dat is ontwikkeld voor het veilig, efficiënt en nauwkeurig anonimiseren van documenten.

#pkdigitaal

Agenda

1. Deelresultaat digitalisering
2. **Deelresultaat zoekingang / interne database**
3. Afsluitende opmerkingen

CBG

#pkdigitaal

PvE en offertetraject

PvE Crowdsource traject en tijdelijk zoeksysteem
Februari / maart 2021
Offertetrajecten april / mei 2021

Leverancier 1	Leverancier 2	Leverancier
Picturae VeleHanden	Het Volk En 42	DeventIT Met Atlantis

terugtrekken

terugtrekken

Kostbare offerte

Matige offerte met weinig diepgang

CBG

#pkdigitaal

Nader onderzoek
mogelijkheden en beperkingen slim OCR/HTR-en

Hoe kunnen we digitale persoonskaarten zo snel mogelijk, zo veel mogelijk geautomatiseerd, transcriberen naar betrouwbare metadata?

Pilot met Multiscan (juni – 23 augustus 2021)
moet uitwijzen:

- wat is de hoogst haalbare metadata kwaliteit is? (de grens van wat kan opzoeken)
- hoe snel is dit te realiseren?
- tegen welke kosten?

Batch 1
7 dozen met persoonskaarten

CBG

#pkdigitaal

Nader onderzoek
mogelijkheden en beperkingen slim OCR/HTR-en

Hoe kunnen we digitale persoonskaarten zo snel mogelijk, zo veel mogelijk geautomatiseerd, transcriberen naar betrouwbare metadata?

Multiscan heeft een partner Scalehub in de arm genomen voor deze pilot:

begin 2016 opgericht
een internationaal multicultureel bedrijf, actief over de hele wereld
uitgebreide ervaring op het gebied van data automatisering, bedrijfsprocesbeheer en crowdsource technologieën, is ingezet om de traditionele data-extractie en documentverwerking revolutionair te veranderen
.
combineert het beste van menselijke vaardigheden en vindingrijkheid met bewezen automatiseringstools.
gebruikt de meest up-to-date methoden voor het converteren van beeldbestanden naar bruikbare gegevens met hoge snelheid en nauwkeurigheid
scherpe gegevensextractie en expertise voor het verwerken van grote volumes

CBG

JT1

#pkdigitaal

Analyse en prioritering metadata Persoonskaart

Prio 1	Prio 2	Prio 3
Familienaam	Beroep	Bijzondere vermeldingen
Voornamen	Geboortedatum echtgeno(o)t(e)	Religie
Geboortedatum	Geboorteplaats echtgeno(o)t(e)	Aantekeningen
Geboorteplaats	Achternaam kinderen	Datum en plaats huwelijk gesloten (idem ontbinding)
Familienaam echtgeno(o)t(e)	Voornamen kinderen	A nummer (voorloper bsn nummer)
Voornamen echtgeno(o)t(e)	Geboortedatum kinderen	Registratie woonadressen
Overlijdensdatum	Geboorteplaats kinderen	Etc.
Overlijdensplaats	Relatie van het kind tot het gezinshoofd, bijvoorbeeldz (zoon), d (dochter), sz (stiefzoon) ofsd (stiefdochter)	
Periode	Achternaam en voornamen ouders	

Scope pilot: focus prio 1, maar ook prio 2 en 3 i.v.m. onderkennen laaghangend fruit

CBG

Dia 16

JT1 Jan Thielen; 31-8-2021

#pkdigitaal

Pilotresultaten: automatisch herkennen typen persoonskaarten

Model kaart	Aantal kaarten	Percentage
Persoonskaarten	5.288	92,5%
Model A	*4.603*	*80,5%*
Model B	*685*	*12%*
Verwijskaart	*129*	2,3%
Onbekend*	297	5,2%
Totaal	**5.714**	**100%**

*Kaarten zijn niet automatisch geclassificeerd. Dit zal handmatig uitgevoerd worden.

Nader onderzoeken

CBG

Pilotresultaten: OCR kwaliteit in %

#pkdigitaal

Prioriteit 1

Veld	OCR-kwaliteit in %
3a – Familienaam	99,00
3b – Voornamen	92,00
4 – Geboortedatum	89,00
4 - Geboorteplaats	97,00
9 – Familienaam echtgeno(o)t(e)	89,00
9 – Voornamen echtgeno(o)t(e)	89,00
Overlijdensdatum*	84,00
Overlijdensplaats*	92,00

*Kwaliteit gebaseerd op die kaarten waarin deze gegevens zijn herkend.
*Percentage van kaarten waarop deze gegevens zijn herkend. Van 5775 zijn er 2317 kaarten waar geen stempel is herkend. Dit betekent niet dat er geen stempel aanwezig is.

Prioriteit 2

Veld	OCR-kwaliteit
7 - Beroep	99,60
11 - Geboortedatum echtgeno(o)t(e)	79,80
11 -Geboorteplaats echtgeno(o)t(e)	99,80
28 - Achternaam kinderen	80,50
28 - Voornamen kinderen	81,00
30 - Geboortedatum kinderen	96,00
30 - Geboorteplaats kinderen	89,00
32 - Relatie van het kind tot het gezinshoofd	98,00
8 - Achternaam en voornamen ouders	99,85
A-nummer (voorloper BSN-nummer)	Niet bekend

Prioriteit 3

Veld	OCR-kwaliteit
23,24,25, 35 - Bijzondere vermeldingen	Niet te definiëren, te divers om hier een percentage aan te hangen
6 - Religie	99,50
33,34 - Aantekeningen	Niet te definiëren, te divers om hier een percentage aan te hangen
13, 14 - Datum en plaats huwelijk gesloten/ ontbinding	96,00
22 - Registratie woonadressen	70,00

In september 2021 controleert het CBG de pilotkwaliteit van de metadata

CBG

#pkdigitaal

Pilotresultaten: wat is kwaliteit?

ABSOLUUT

Aantal karakters goed vertaald, uitgedrukt in % van totaal aantal karakters

Voor de prio 1 velden is het streven zo dicht mogelijk bij 100% kwaliteit.
Streven naar 99,9%?
(nog af te spreken)

RELATIEF

Kwaliteit die goed genoeg is voor het herkennen van de juiste persoonskaart:

3a - Familienaam	Herkende familienaam
Aalfs	aAalfs
Aalderink	X Aalderink
Aarntzen	A_Aarntzen

CBG

#pkdigitaal

Agenda

1. Deelresultaat digitalisering
2. Deelresultaat zoekingang / interne database
3. **Afsluitende opmerkingen**

CBG

#pkdigitaal

Van ambitie naar gefaseerde realisatie

Kortom

De pilotresultaten geven veel stof tot nadenken. De maand september 2021 gebruiken we voor het analyseren van de pilotresultaten en de vertaling naar een best haalbare planning die tevens rekening houdt met de beschikbare middelen.

#pkdigitaal

Van ambitie naar gefaseerde realisatie

Fase 1: planning en ambities
Digitaliseren: 5,8 miljoen PK's in 8 maanden digitaliseren (januari 2022)
Masterbestanden duurzaam offline bewaren in een Azure netwerk
Afgeleide (JPEG) bestanden online beschikbaar voor verbetering huidige dienstverlening (efficiënter **richting kostendekkend** en hogere beeldkwaliteit voor de klant)
OCR / HTR 5-10% metadata (prio 1) + eerste versie tijdelijk zoeksysteem (maart 2022)
Fase 2: ambities
OCR / HTR stapsgewijs naar 100% metadata (10%, 20%, 33%, 50%, 100% prio 1) + verbeterd tijdelijk zoeksysteem (december 2022)
Experimenteren met automatische matching (juli – december 2022)
Fase 3: ambities
Toekomstvast NRO inclusief automatische matching (voorjaar 2023)
Keuzes maken: prio 2 en 3 velden metadateren

Dekking dankzij projectbijdrage vanuit Ministerie van Binnenlandse Zaken

Ambities vertalen naar planning, nieuwe subsidieaanvraag

#pkdigitaal

Onzekere factor: corona invloed