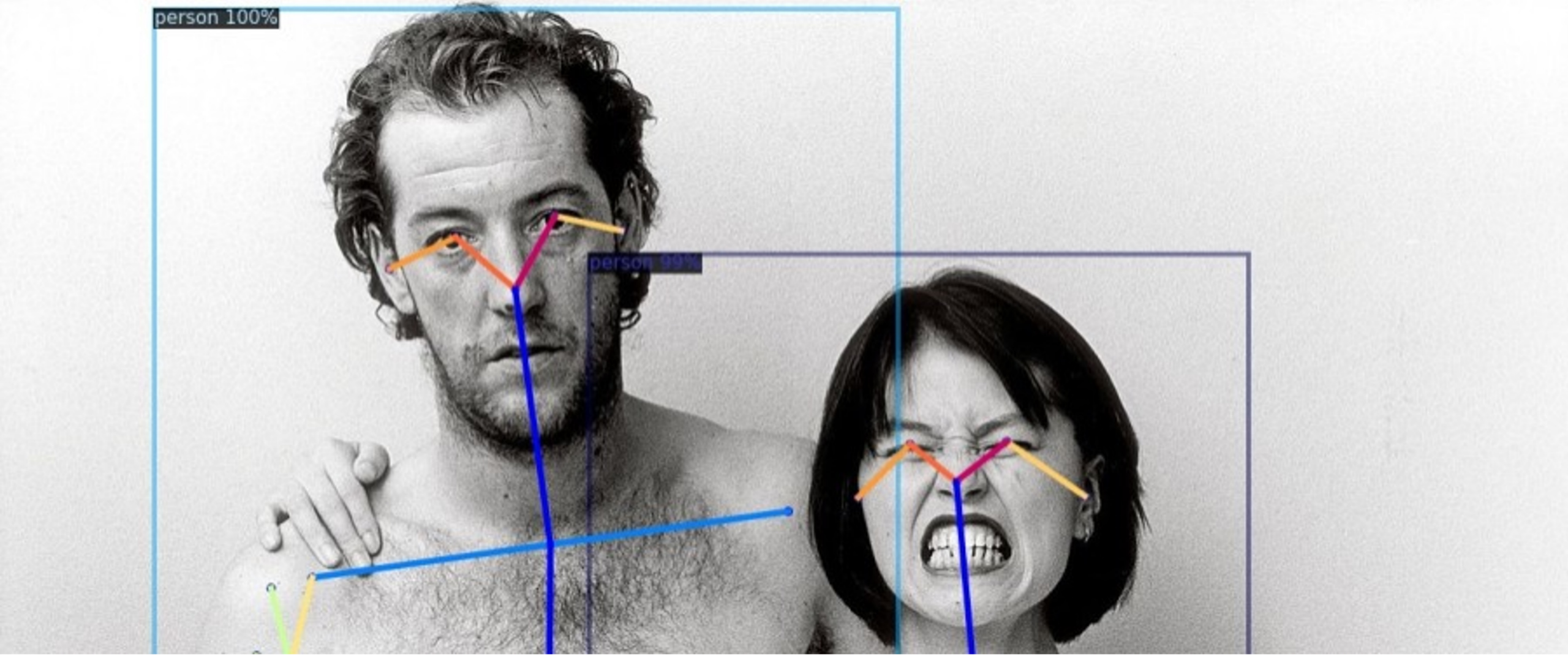


GiVE metadata

Stand van zaken - collegagroep - 16 maart 2023

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

GiVE metadata - Context

- Grote hoeveelheden digitaal materiaal gearchiveerd
  - digitaliseringsprojecten afgelopen jaren
  - digitaal geboren content
- Manuele metadering is duur en tijdrovend
- Tegelijkertijd: tooling op basis van AI wordt matuur

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Relance middelen - GiVE metadata

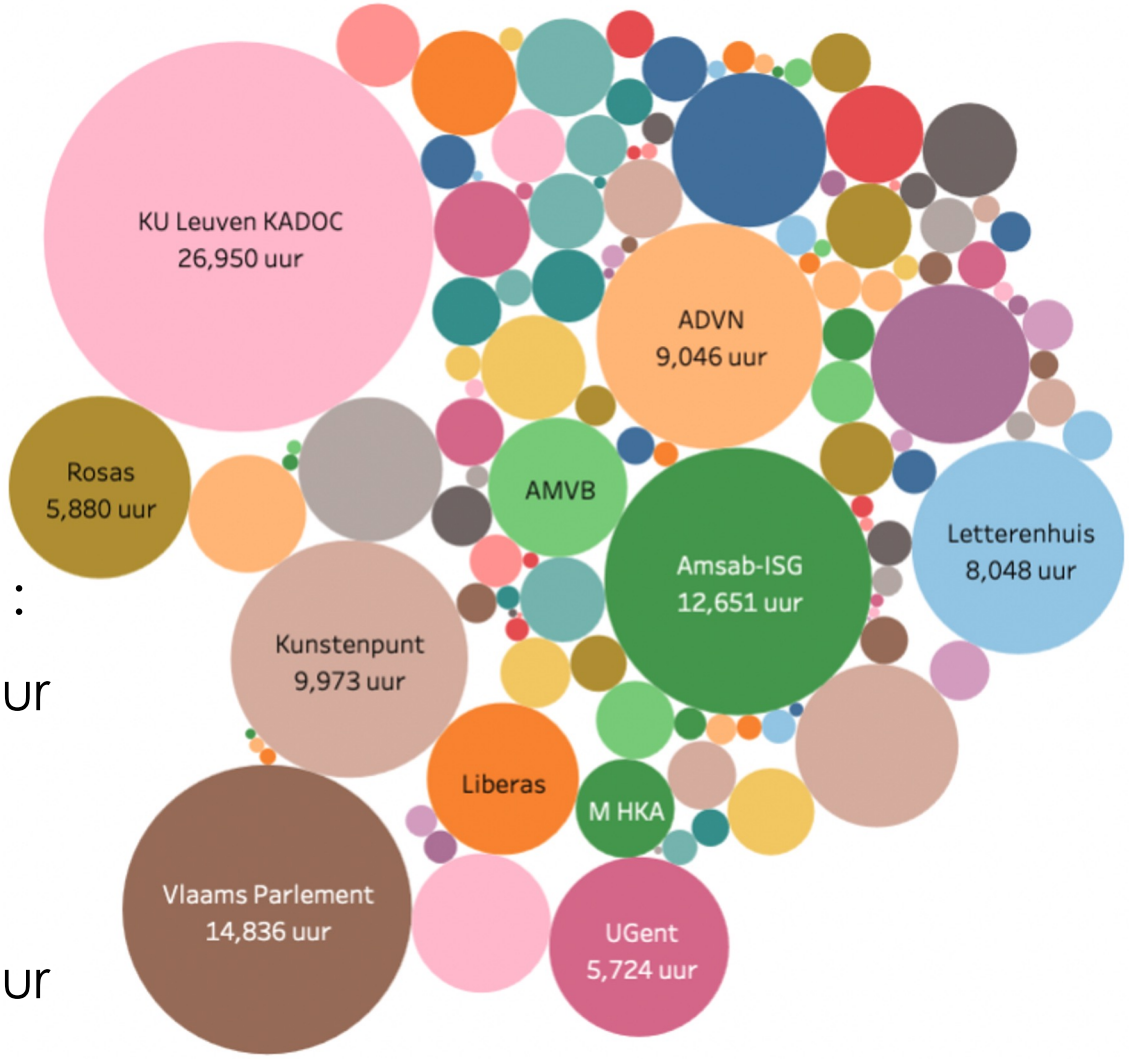
- Wat willen we doen?
  - **Spraakherkenning** op Nederlandse spraak (audio/video)
  - **Entiteit herkenning** op tekst (Personen, locaties)
  - **Gezichtsdetectie** op video ; gezichts**herkenning** op een beperkte set personen

- Op welke collecties?
  - Alle reeds gearchiveerde AV-collecties (behalve die van omroepen)

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Schaal

- 120 contentpartners betrokken
- Per activiteit
  - Spraakherkenning en Entiteit Herkenning :
    130.000 gearchiveerde stuks of 160.000 uur media
  - Gezichtsdetectie en -herkenning :
    100.000 gearchiveerde stuks of 120.000 uur media



meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Verdere projectgegevens

- Timing: najaar 2021 tot eind 2023
- Workflows die later ook herbruikbaar zijn
- Privacy en ethische aspecten
- Mature technologie, beperkte onderzoekscomponent
  - We werken verder o.b.v. resultaten FAME en eerder onderzoek in de media en CE sectoren.
  - Nauw contact met organisaties (bv. VRT, Beeld en Geluid, BBC, ..) die hier ervaring mee hebben.

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Veel partners

- We werken met materiaal van heel veel partners
- Vaak ook nieuwe uitdagingen / inzichten
- Informatie via communicatieplan
- Betrokkenheid via werkgroep
  - krijgen meer in-depth informatie
  - bepalen mee wat we doen en niet doen
    - bvb. parametrisering gezichtsherkenning
    - bvb. beheer referentieset gezichten.

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Deze presentatie

- Tussentijdse stand van zaken
  - Work in progress
  - Blik in de keuken
- Juridische aspect
- Spraakherkenning
- Gezichtsherkenning
- Future work

Beeld: De keuken, Louis Thevenet ; Collectie museum Dhondt-Dhaenens Fotograaf: Cedric Verhelst

public domain

Juridische aspecten

- AI toepassen: het kan volgens GDPR (archivering in het algemeen belang)

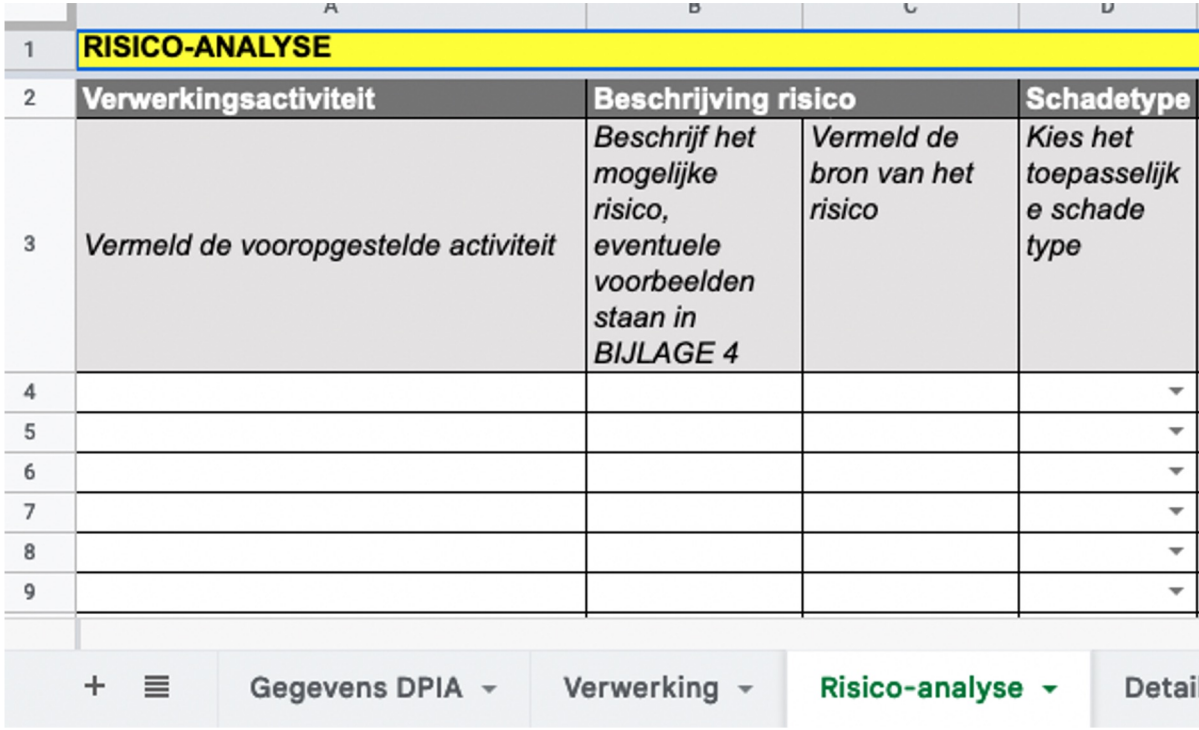
- Maar DPIA is nodig want:
  - Grootschalige verwerking
    - → Aantal betrokkenen
    - → Volume van de gegevens
    - → Duur van de activiteit
  - Creatie van nieuwe metadata kan linken leggen tussen personen en lidmaatschap vakbond/etniciteit/politieke voorkeur…

    → Verwerking van 'bijzondere categorieën persoonsgegevens'

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Data Protection Impact Assessment

- Deel 1 : omschrijf wat je wil doen
  - Algemene beschrijving beoogde verwerking
  - Beschrijving type persoonsgegevens
  - Doel van de verwerking
  - Bronnen van de persoonsgegevens
  - Betrokkenen
- Deel 2 : Risico analyse
  - Wat zijn de taken in het project?
  - Welke risico's zijn hieraan verbonden?
  - Hoe gaan we deze minimaliseren?



	A	B	C	D
1	**RISICO-ANALYSE**			
2	**Verwerkingsactiviteit**	**Beschrijving risico**		**Schadetype**
3	*Vermeld de vooropgestelde activiteit*	*Beschrijf het mogelijke risico, eventuele voorbeelden staan in BIJLAGE 4*	*Vermeld de bron van het risico*	*Kies het toepasselijke schade type*
4				
5				
6				
7				
8				
9				

Gegevens DPIA | Verwerking | Risico-analyse | Detai

⇒ Checklist voor privacy aspecten binnen project

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Ethische aspecten

- ism. Kenniscentrum data & maatschappij
- Meerdere workshops, focus op gezichtsherkenning
  - breng alle stakeholders samen
    - archivarissen, personen die herkend zullen worden, technici
  - Probeer tot een principes document te komen of gedeeld inzicht / proces
  - Bvb. referentielijst

Ethische / juridische conclusies worden samengevat & gedeeld

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Spraakherkenning

- Relatief mature producten in de markt
- Marktbevraging eerste helft 2022
  - Wat is mogelijk op dit moment, hoe snel kan de verwerking (snel!), …
  - Informele gesprekken die ons inzicht leverden in wat kan
- Europese aanbestedingsprocedure tweede helft 2022
- Criteria
  - Prijs (om 1 uur te transcriberen)
  - Kwaliteit via benchmarking (zie volgende slides)
- 5 deelnemers:
  - Azure, **Speechmatics**, Scriptix, Amberscript, Notubiz

Benchmark - objectieve kwaliteitsmeting

Dataset

- Handgeselecteerd uit het archief
- 5 hoofdcategorieën: Radio/TV interview, Politiek debat/interview, Spontaan commentaar (sport/event), Reportage/Documentaire, Nieuwsbulletin
- Nevencategorieën: podiumkunsten, dialect, oud materiaal, andere taal
- 165 bestanden, >3 uur audio

Data Annotatie (extern bureau)

- Ground Truth transcripties (letterlijk)
- Keyword annotaties: locatie, persoon, organisatie, belangrijke kernwoorden

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

STT Benchmark - methodologie

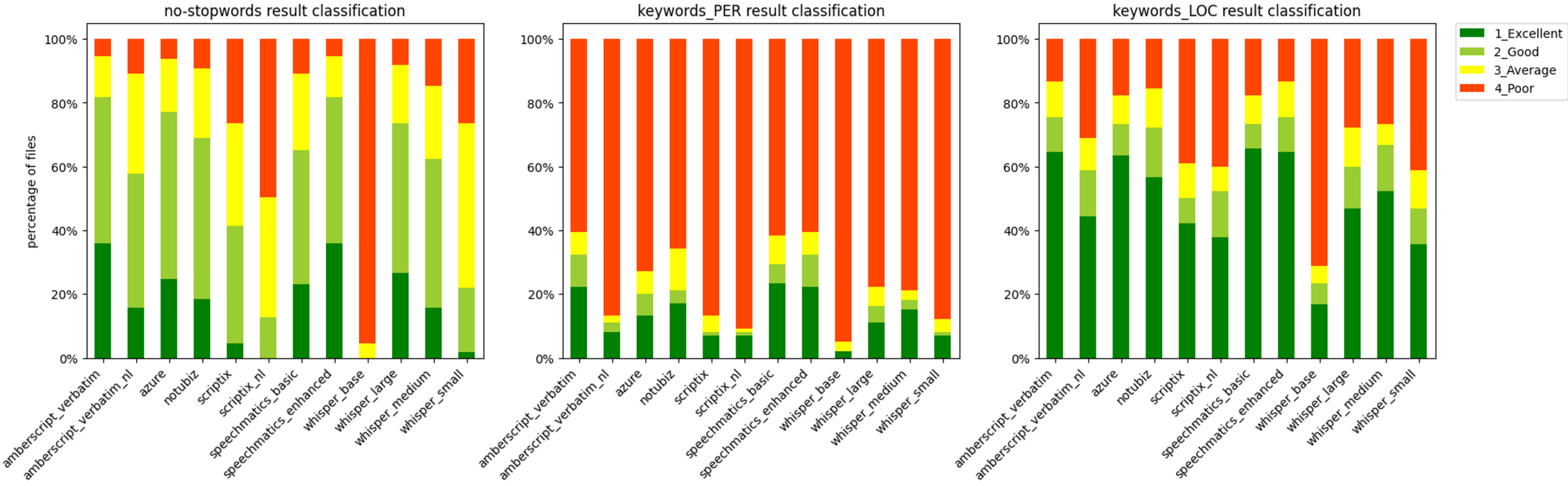
STT solutions

- SaaS: Speechmatics, *Amberscript*, Notubiz, Scriptix, Azure
- Whisper: OpenSource multi-language STT model (Open AI)

Benchmark tool

- Gebaseerd op EBU benchmark-stt
- WER (Word Error Rate): hoe lager, hoe minder fouten in de transcriptie
- Verschillende WER scenarios: no-stopwords, keywords (PER, LOC, ORG, KEY, ALL)
- 4 kwaliteitsklasses: Excellent (0-0.1), Good (0.1, 0.2), Average (0.2-0.3), Poor (>0.3)
- Totale kwaliteitsscore: gebaseerd op hoeveel files onder welke kwaliteitsklasse vallen
- Enkel kwaliteit van hoofdcategorieën telt mee voor kwaliteitsscore

STT Benchmark - results



Conclusie: Speechmatics heeft de beste transcriptiekwaliteit

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Gezichtsherkenning - Wat willen we?

- Gezichten **identificeren** en **herkennen** in video (> 100k uur)
  - Grote volumes gezichten, grote volume aan data
- Referentieset: te herkennen gezichten
  - Hoe gaan we die samenstellen?
  - Hoe beheren?
  - Gedeelde referentieset?
- Vaak voorkomende gezichten, niet gelinkt aan referentieset
  - Kunnen we hier rond functionaliteit uitbouwen?
  - Bvb. top X meest voorkomende gezichten in je collectie
  - Opportuniteit om referentieset uit te breiden.

Gezichtsherkenning - Aanpak

- Analyse
  - Kopen?
    - Marktbevraging
    - Wat kan op dit moment?
    - Wat is de kost?
  - Bouwen
    - Kunnen we verder op FAME bouwen?
    - Wat is de kost?
- Ethische aspecten & betrokkenheid gebruikers
  - Functionele analyse ism. werkgroep
  - Kenniscentrum data & maatschappij

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Gezichtsherkenning - kopen vs bouwen

- Kopen
  - Marktbevraging
    - AWS
    - Azure
    - Vicarvision
  - Kosten
    - Operationele kost
    - Relatief duur
  - Meer geavanceerde cases zijn moeilijker te realiseren
  - Privacy & ethiek
- Bouwen
  - Meer vrijheden
  - Meer op maat van onze use cases / content partners
  - Technische uitdagingen
    - Kan het überhaupt?
    - Welke modellen zijn nodig?
    - Zijn ze open source?
  - Wat zou de kost zijn om dit te bouwen?

Uiteindelijk gekozen om dit te bouwen, vertrekkende van FAME

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Gezichtsherkenning - bouwen op FAME

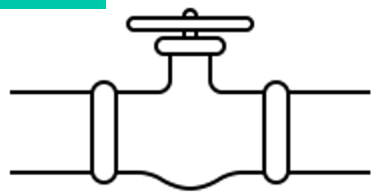
- FAME
  - Gezichtsherkenning op vnl. foto.
  - Reeds uitgebreide referentieset samengesteld
  - Basisflow voor detectie en herkenning uitgewerkt
  - Heel wat privacy aspecten onderzocht

- Uitdagingen
  - Schaal (foto vs. video)
  - Matching algoritme helemaal anders (geen validatie)
  - Gedeelde referentieset vs referentieset per partner
  - GiVE = geen research (bvb. gebruik modellen zoals insightface)

Bouw pipeline

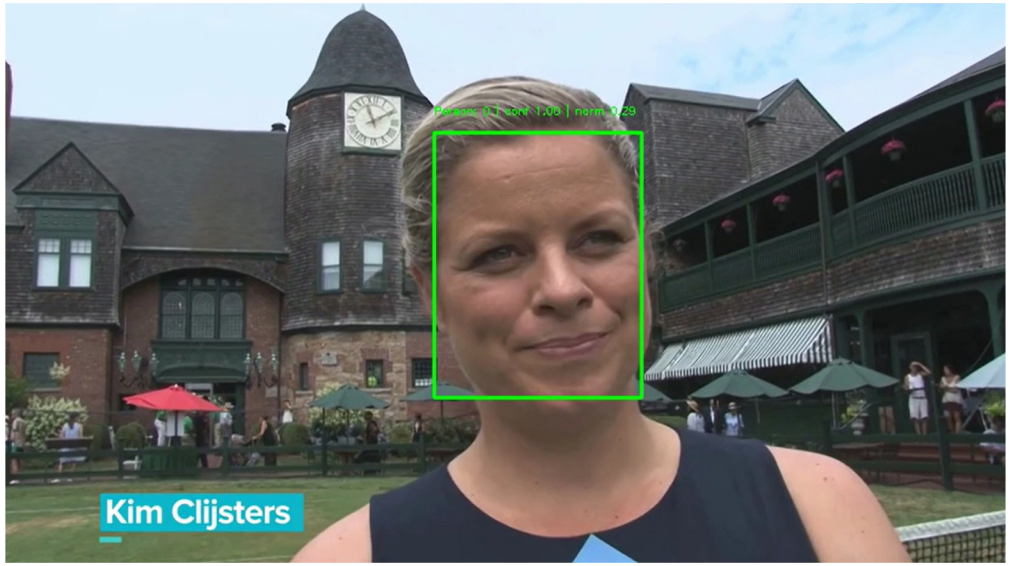
- Team externe consultants (Cronos, IT Planet, Ordina)
  - specialisten AI + Machine Learning
  - programmeur workflows
  - functionele analist (referentie set beheer)
  - project management
- Samenwerking met werkgroep
  - Sessies rond parametrisering
  - Uitgebreide functionele analyse: wat willen jullie?

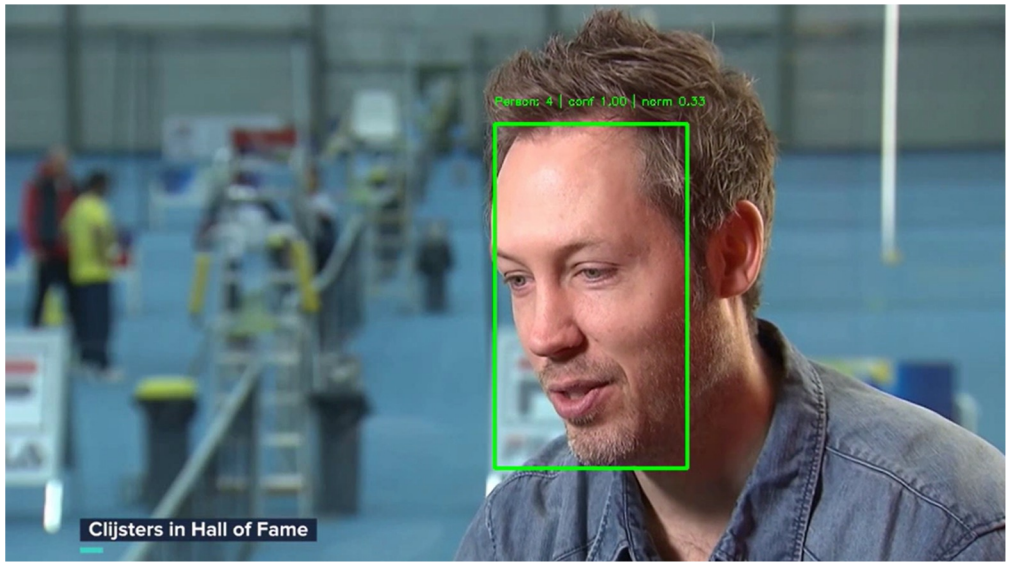
Gezichtsherkenning - Video Pipeline



Video pipeline

Face Detection → Face Embedding → Face Tracking → Face Clustering → Face Identification

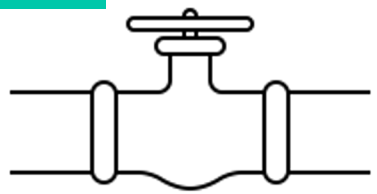




Detecteer gezichten in 1 frame

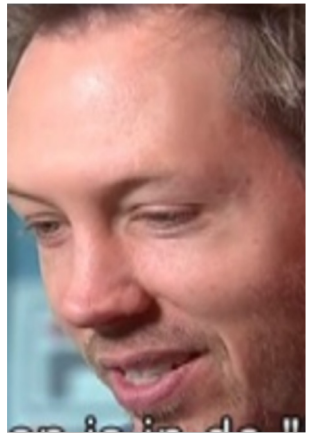
meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Gezichtsherkenning - Video Pipeline



Video pipeline

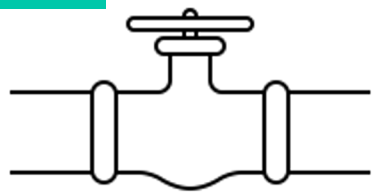
Face Detection → **Face Embedding** → Face Tracking → Face Clustering → Face Identification



Bereken fingerprint van elke face

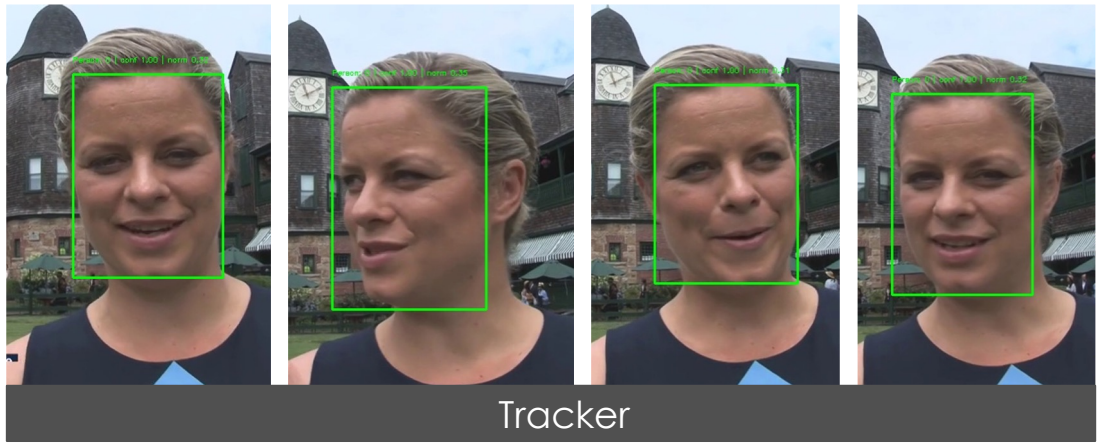
meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

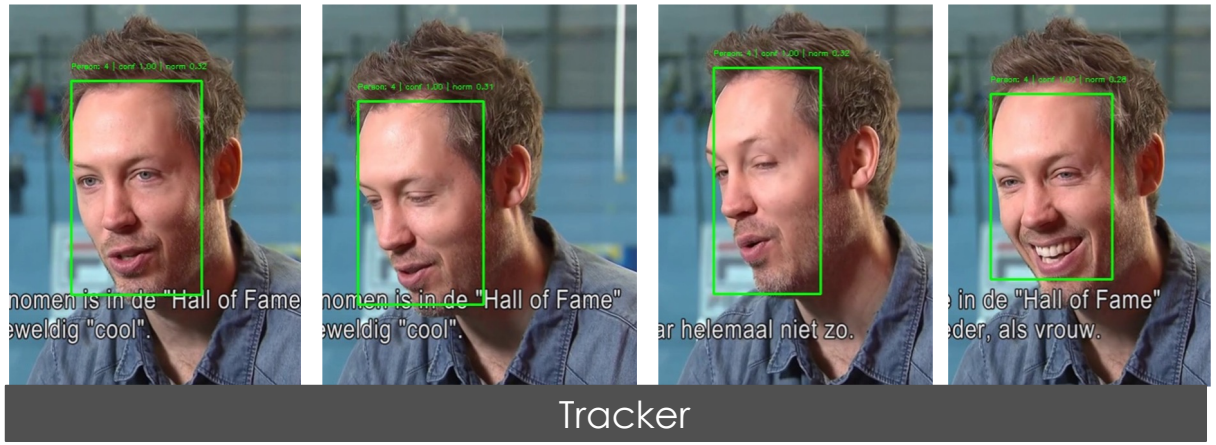
Gezichtsherkenning - Video Pipeline



Video pipeline

Face Detection → Face Embedding → **Face Tracking** → Face Clustering → Face Identification

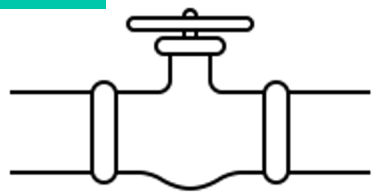




Volg een gezicht van een persoon over meerdere frames binnen 1 shot tot trackers

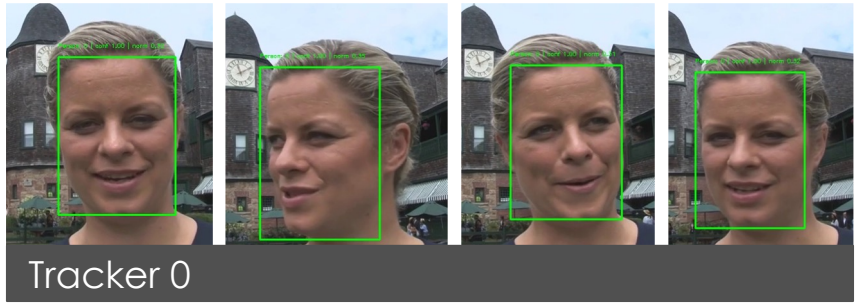
meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

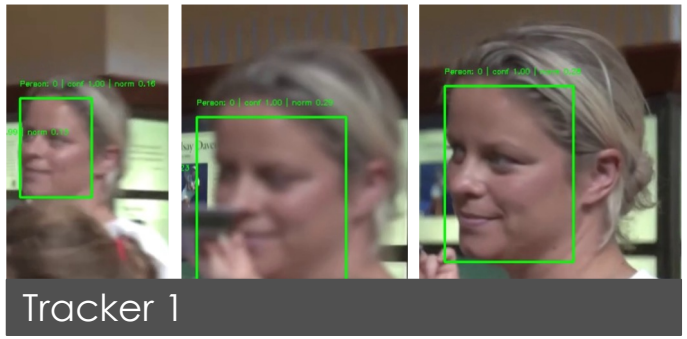
Gezichtsherkenning - Video Pipeline



Video pipeline

Face Detection → Face Embedding → Face Tracking → **Face Clustering** → Face Identification

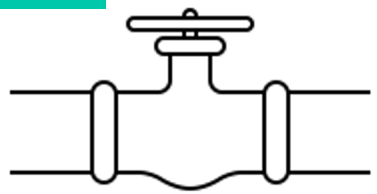




Cluster groepen van gevolgde gezichten (trackers) bij elkaar tot personen

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Gezichtsherkenning - Video Pipeline

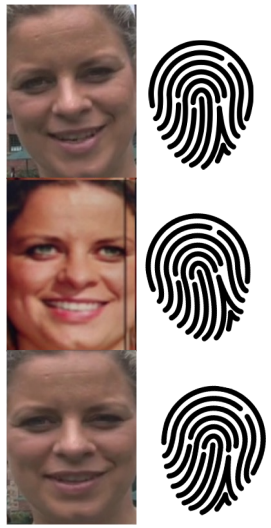


Video pipeline

Face Detection ▶ Face Embedding ▶ Face Tracking ▶ Face Clustering ▶ Face Identification

Faces Persoon 0

Subset obv kwaliteit



- Goede gelijkenis
- Zelfde persoon

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Gezichtsherkenning - parameters

Gezichtskwaliteit

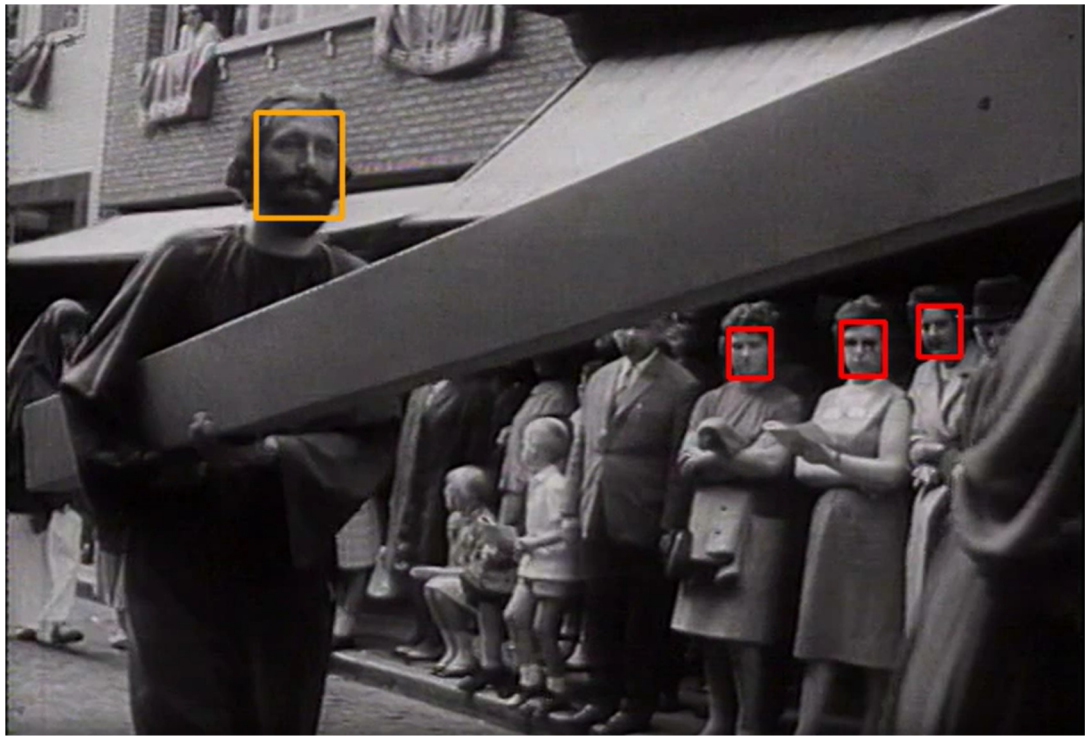
- Box-grootte
- Pose (landmarks)
- Belichting
- Blurriness
- Confidence

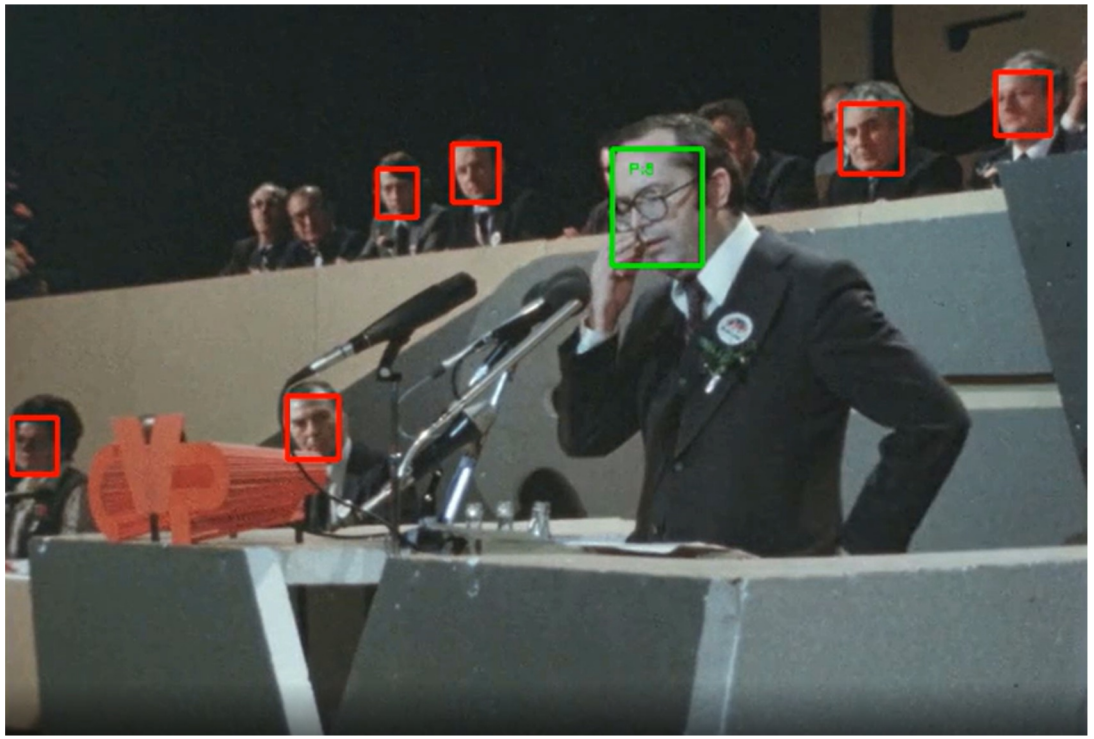
Schermtijd

- Duur van gezicht binnen 1 shot (tracker)
- Aantal keer dat gezicht terugkomt (aantal trackers per persoon)
- Totale schermtijd van 1 persoon in video

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Nu: parametrisering met werkgroep





Oranje gezichten komen te kort voor (bv. kort shot binnen reportage); gezichten in de achtergrond leveren vaak geen metadata op omdat ze te klein zijn

05:48
http://give-face-gas.private.cloud.meemoo.be/task_results/baba7b5ef1d2459180c2ea8fffa49d02cf2aed7b29ba41758f43c7e1eae457a1b6a6377d6b4d4e2c8b032c69a4597bbc_af862059c93c6475d5320919c7c48b7d

2:00 http://give-face-gas.private.cloud.meemoo.be/task_results/99496d03b4904e1fa23c6b0a203d36db8651dc32f64049bb995ffd93dbc395c28cb9a1f48db84a5c9b10192aba6e9d2c_af862059c93c6475d5320919c7c48b7d

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

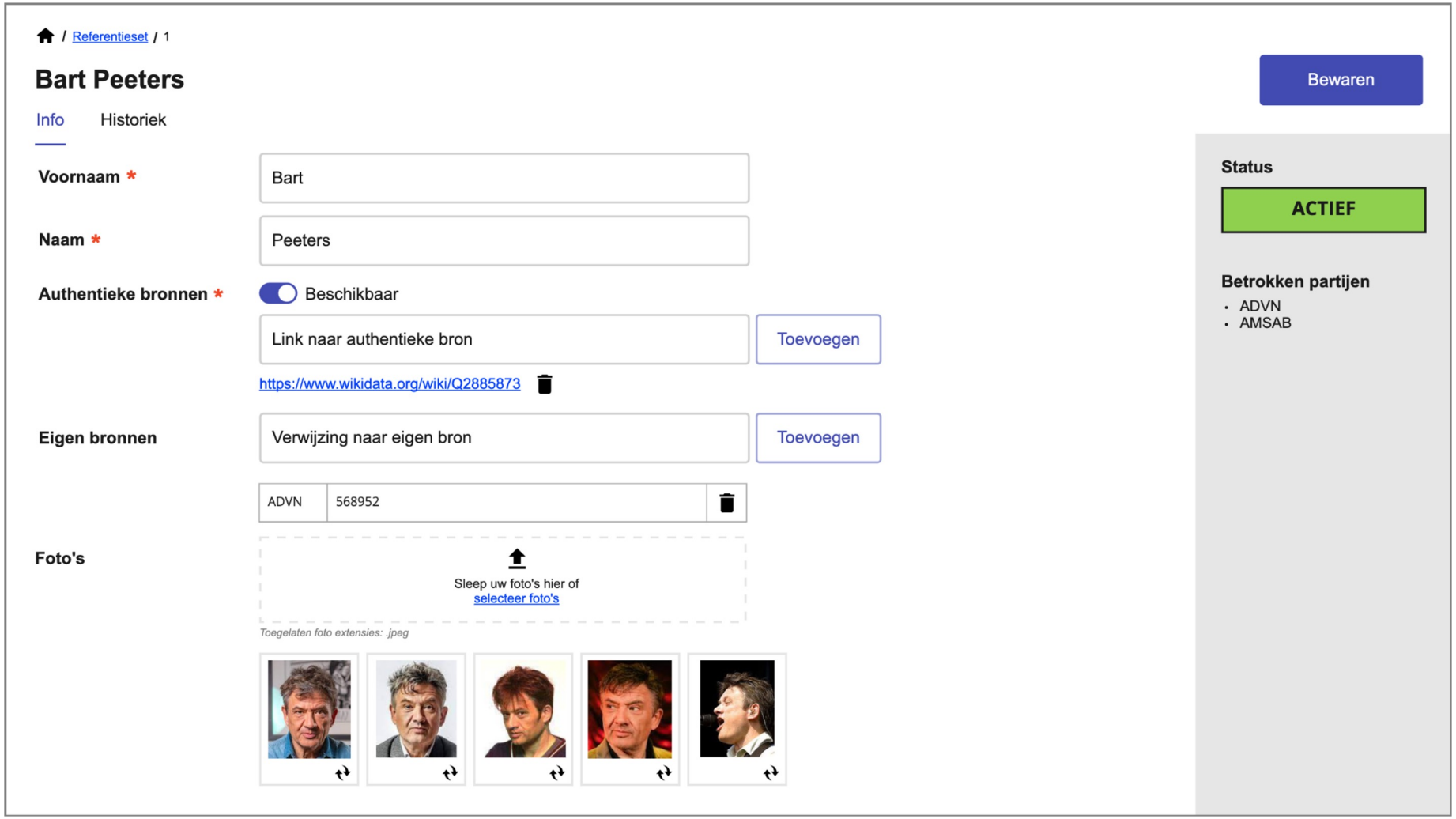
Referentieset en gebruik van tooling

- Basis: FAME referentieset
  - Foto's te herkennen personen
  - Identifiers met link naar publieke bronnen waar mogelijk
- Interviews + sessie werkgroep leden
  - Wat moeten we kunnen in het beheer van de ref set?
  - Wat willen we precies kunnen?
- Doel: gedeelde visie
  - Achterliggende processen
  - Daaruit: vereisten voor de software + wireframes

Referentieset - voorlopige resultaten

- Gedeeld beheer door CP's
  - Geen ownership van de data
  - Gebruikers kunnen alle referentie set entries aanpassen
  - Wel gedetailleerd inzicht in historiek en aanpassingen
- Links naar zowel publieke als private bronnen
  - Bvb. wikidata
  - Bvb. interne identifiers
- Begeleiding van de beheerders
  - Richtlijnen / checks op fotokwaliteit

Gezichtsherkenning - Functionele analyse



meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Gezichtsherkenning - Functionele analyse

/ Referentieset / 1

Bart Peeters

Bewaren

Info Historiek

5 januari 2023 11h03	Foto's toegevoegd 	Phaedra Claeys (ADVN)
3 januari 2023 15h23	Eigen bron '568952' toegevoegd	Phaedra Claeys (ADVN)
1 januari 2023 10h02	Nieuwe persoon 'Bart Peeters' aangemaakt	Kim Robensyn (AMSAB)

Status

ACTIEF

Betrokken partijen

- ADVN
- AMSAB

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

(Meta) Metadata

Metadata uit machine learning is dynamisch (processen verbteren, nieuwe referentiepersonen, etc.)

Welke "**provenance**" data en **historiek** biedt **meerwaarde** voor content partners ?

- aanmaak (metadata door AI of manueel aangemaakt)
  - datum
  - indien manueel: naam van persoon, organisatie
  - specifieke AI meta
    - Spraak: API version
    - Gezicht: model version
- versioning & granulariteit per update
  - Gezicht:
    - herkende personen toegevoegd aan metadata file ?

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

Verder werk

- NER - analyse loopt op dit moment
- Voorjaar 2023
  - Opstart pipeline spraakherkenning
  - Opstart pipeline gezichtsdetectie
  - Opstart pipeline NER
- Eind 2023
  - Ter beschikking stellen resultaten aan partners.
  - Finale resultaten / deliverables beschikbaar

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF

EFRO
EUROPEES FONDS
VOOR REGIONALE
ONTWIKKELING

Europese Unie

Dit project kadert binnen het relanceplan Vlaamse Veerkracht en wordt gerealiseerd met de steun van het Europees Fonds voor Regionale Ontwikkeling.

meemoo
VLAAMS INSTITUUT VOOR HET ARCHIEF