

AI in erfgoedcollecties

Een perspectief uit Vlaanderen...



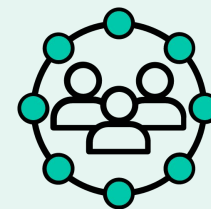
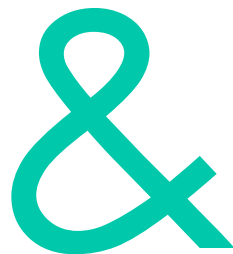
Bart Magnus - 16/06/2026



Twée strategische hoofdlijnen

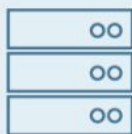


Samen met onze partners
duurzaam bouwen aan het
digitale geheugen
van Vlaanderen



Mensen en organisaties actief
doen **participeren** in
en interageren met (digitaal)
erfgoed

In cijfers



Objecten duurzaam
bewaard in ons
archiefsysteem >

8.258.546



Gedigitaliseerde en
digitaal gemigreerde
audiovisuele dragers >

550.101



Contentpartners
waarmee wij
samenwerken >

184



Vragen om advies
(loketvragen)
beantwoord in 2024 >

260



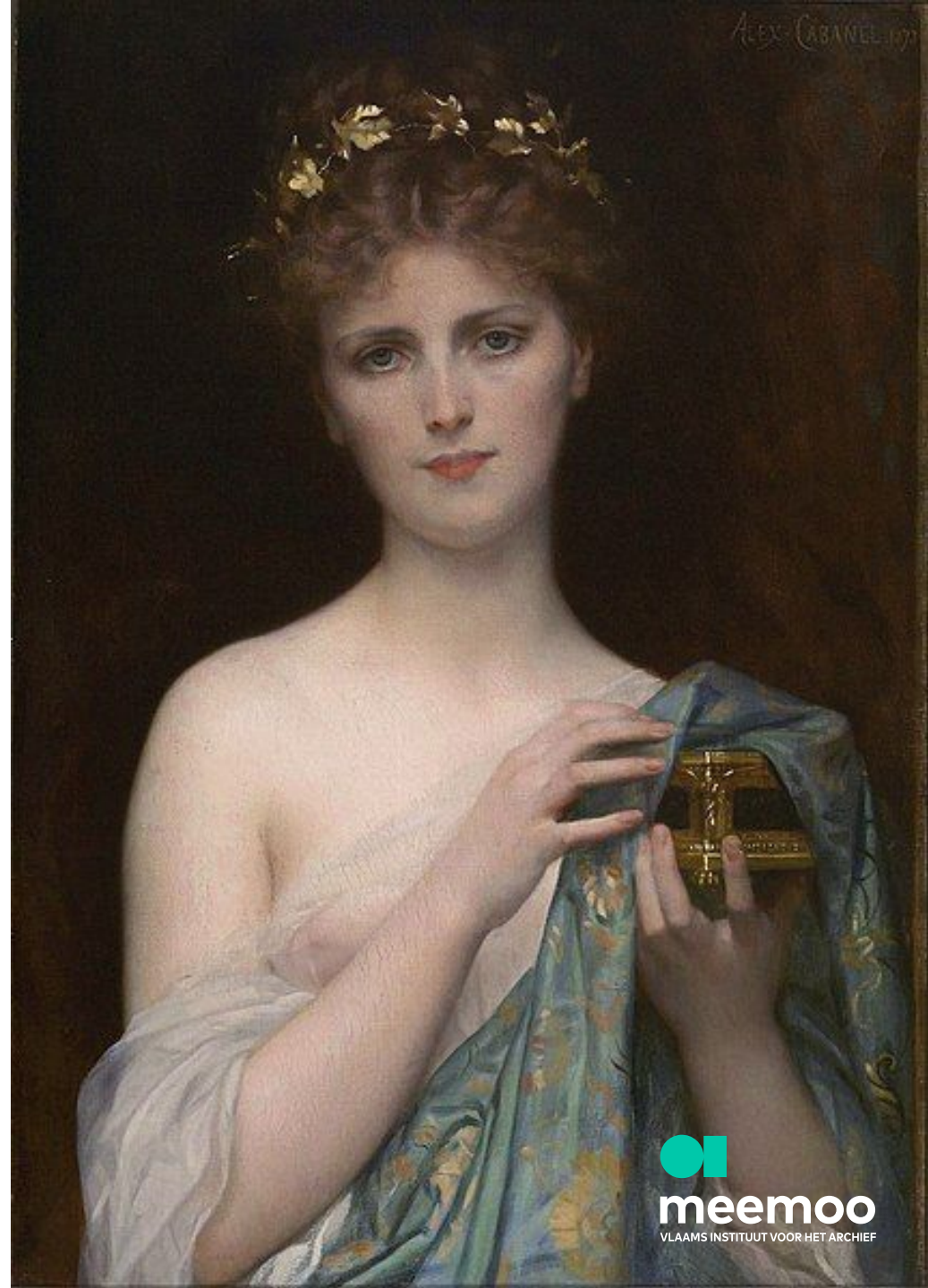
Events en vormingen
voor media, cultuur en
overheid in 2024 >

30

AI in erfgoedcollecties

welke meerwaarde?

Pandora, door Alexandre Cabanel (1873), collectie: Walters Art Museum. Publiek domein, via Wikimedia Commons



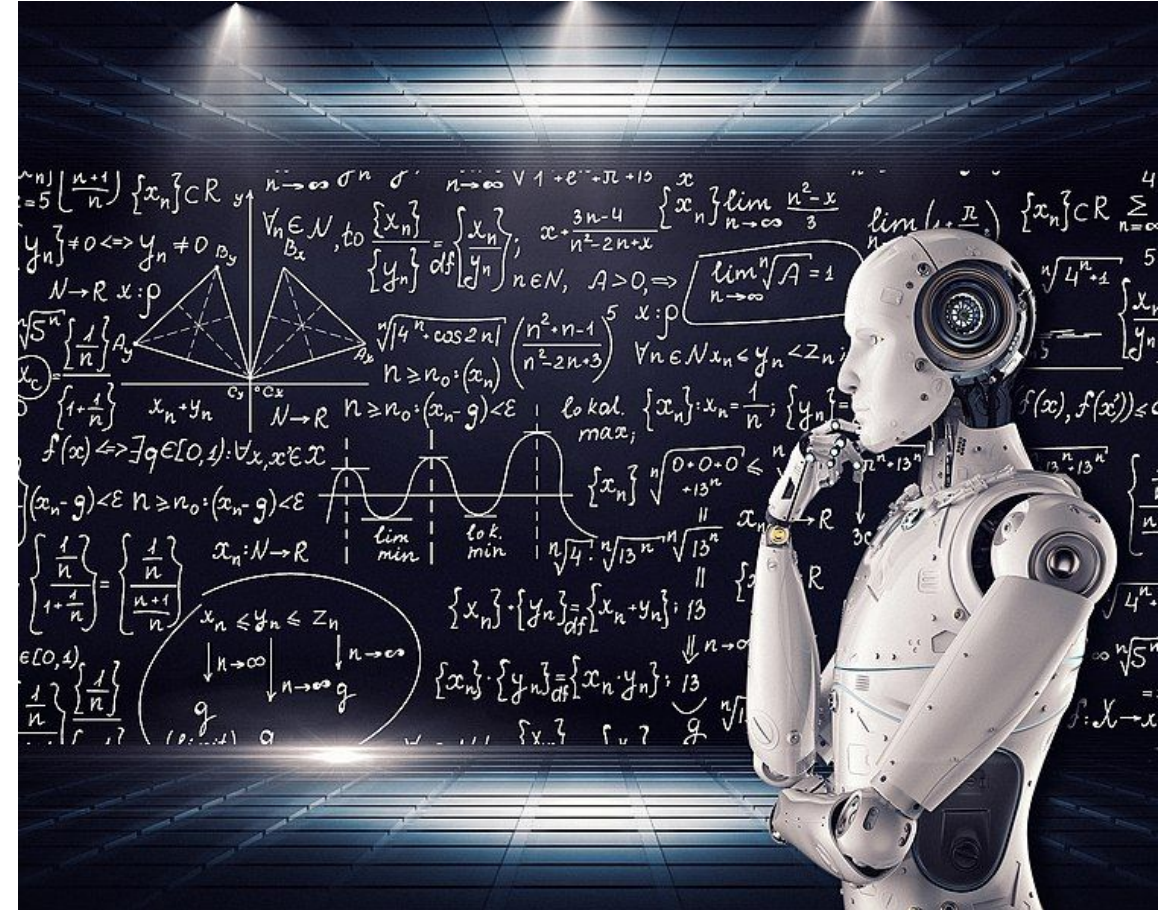
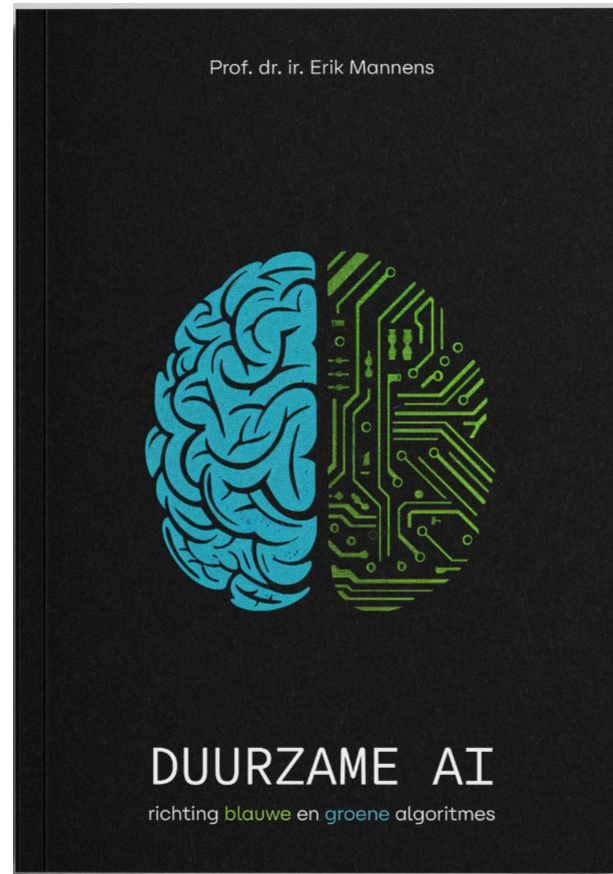
What's in a name?

AI

- artificieel?
- intelligent?

Machine learning

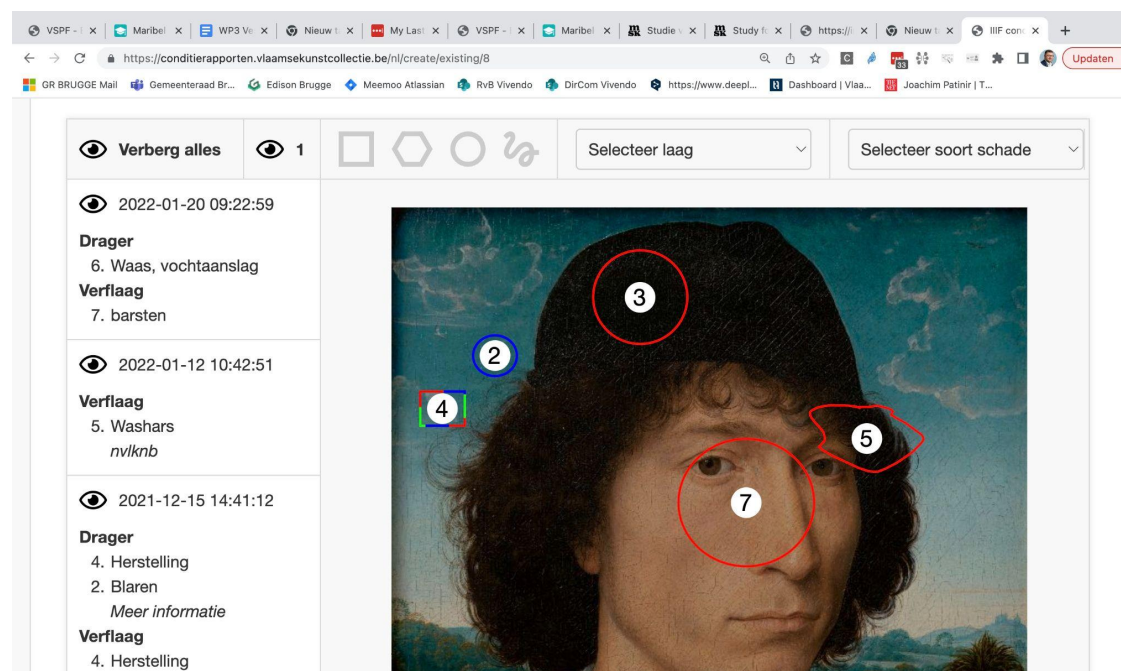
- algoritmes
- wiskunde
- rekenkracht



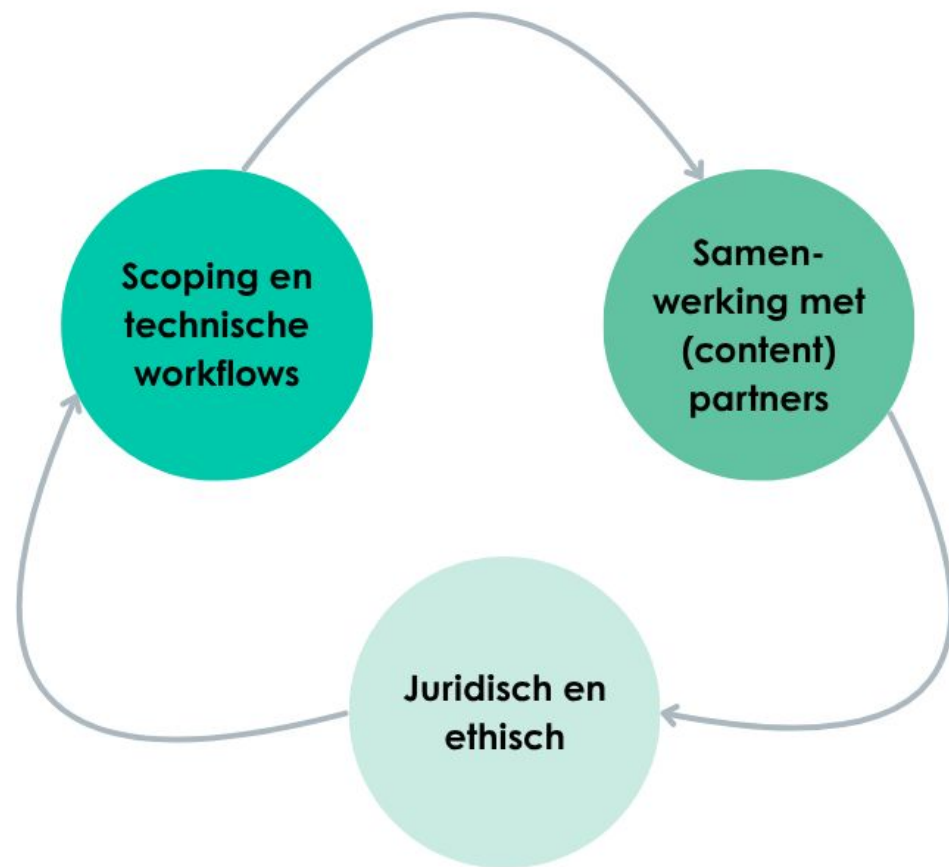
<https://www.vpnrus.com/>, [CC BY 2.0](#), via Wikimedia Commons

Potentieel AI in erfgoedsector

- metadata corrigeren
 - onaangepaste terminologie (bv. koloniaal, racistisch)
- metadata creëren -> doorzoekbaarheid
 - beeldherkenning
 - gezichten
 - locaties
 - gebouwen
 - concepten
 - spraakherkenning
 - OCR/HTR en NER/NEL
- conditierapportage



AI bij meemoo



FAME (2021-2022)

 **KOERS**
MUSEUM VAN DE WIELERSPORT

FAcial recognition for **ME**tadata creation

Waarom?

Weinig goede metadata bij (gedigitaliseerde) foto's en video's belemmert

- online toegang
- vindbaarheid
- hergebruik



FAME: 3 pilootprojecten

- **podiumkunstenaars** (i.s.m. Kunstenpunt)
 - 8.300 foto's van producties, 700 (groeps)portretten + enkele video's
- **wielrenners** (i.s.m. KOERS)
 - 115.000 fotonegatieven uit de periode 1967-2006
- **politici en activisten** (i.s.m. ADVN en Archief van het Vlaams Parlement)
 - ADVN: 8.300 foto's + enkele gedigitaliseerde video's
 - Archief van het Vlaams Parlement: 5.300 foto's + enkele video's van zittingen uit de periodes 1992-1994 en 2001-2021.

⇒ “publieke” personen

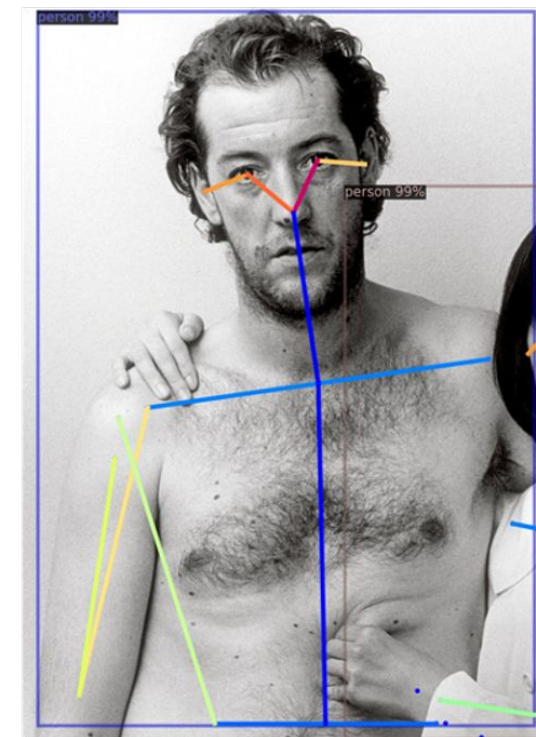
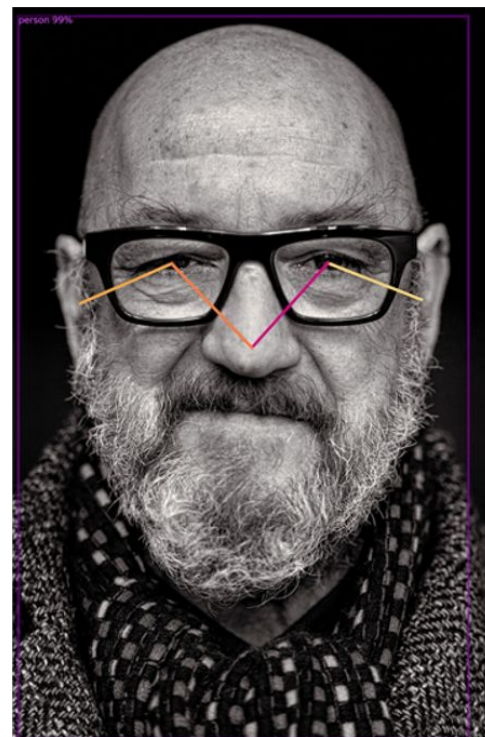
FAME: referentiesets



- Wie willen we herkennen?
- nood aan foto's waarvan we weten dat persoon x er op staat afgebeeld
 - via partners
 - via andere collectiebeherende instellingen
 - semi-automatisch (vb. via API Wikidata)
 - manueel (tijdsintensief)
- **54.540 portretfoto's** van in totaal **6.075 verschillende personen**
 - 2.393 podiumkunstenaars (37.172 referentiefoto's);
 - 2.791 wielrenners (15.323 referentiefoto's);
 - 891 politici (2.045 referentiefoto's)

Manuele validatie

- Doel: zo betrouwbaar mogelijke resultaten > hoe goed/slecht doet het algoritme het?
- Door medewerkers van collectiebeherende organisaties
- online tool
 - gezichtsdetectie
 - similariteitsscore: automatisch herkende gezichten met similariteitsscore van minstens 0,5 zijn gemiddeld 93,6 % betrouwbaar



Relevantie similariteitsscore

Nieuwe metadata terug naar bronsystemen?

- Welke foutenmarge aanvaardbaar?
- Vanaf wanneer is manuele validatie noodzakelijk?



Gezichtsherkenning > ethische aandacht

AI voor metadatacreatie: gezichtsherkenning (2021 - nu)

ethiek als verlengde van juridisch kader:

- auteursrecht
- recht op afbeelding
- AVG (Algemene Verordening Gegevensbescherming)

Ethische aspecten

- **Samenstelling referentiesets**
 - keuze voor publieke figuren beperkt impact op privé-personen;
 - afweging uiteindelijke doel tegenover eventuele negatieve gevolgen voor afgebeelde personen;
 - delen van referentiesets? efficiëntiewinst, maar ook gevaar op verlies aan controle > nood aan beroepsethiek.
- **Vooringenomenheid (*bias*) in algoritmes**
 - zorgvuldig gecreëerde referentiesets;
 - controle van gebruikte algoritme op *bias* aan hand van manuele steekproef: worden alle mensen even goed herkend worden, ongeacht hun fysieke eigenschappen of kledij?

GEZICHTSHERKENNING om erfgoed te metadateren: mag dat?

Gezichtsherkenningstechnologie helpt om personen in fotografische en audiovisuele collecties te herkennen en zo rijkere metadata te genereren. In dit en het volgende nummer bekijken we met een ethische en juridische bril of dat zomaar kan. Deel 1: het ethische luik (uitgebreider besproken in het juninummer van *faro* | tijdschrift over cultureel erfgoed).



REFERENTIESETS

De toepassing van gezichtsherkenningstechnologie vereist goede referentiesets. Dat zijn verzamelingen foto's die met zekerheid een bepaalde persoon afbeelden. Een algoritme vergelijkt die referentiefoto's met de foto's uit de doelcollectie en voorspelt met welke waarschijnlijkheid afgebeelde personen daar ook in voorkomen. Referentiefoto's kunnen uit diverse bronnen komen, waaronder het web. Hoe ethisch is het massaal verzamelen (inclusief scrapen) en bewaren van foto's op het web zonder toestemming van de afgebeelde personen? Is er een verschil wanneer de foto's tot het publiek domein behoren of gepubliceerd zijn onder een vrije licentie? Er ontstaat een spanning tussen open content en ethische overwegingen wanneer open content gebruikt wordt voor gezichtsherkenning, al dan niet door grote commerciële spelers of overheden. Het uiteindelijke doel en mogelijke negatieve gevolgen bepalen of het massaal verzamelen van foto's voor gezichtsherkenning ethisch verantwoord is.

Een stap verder is het opstellen van referentiesets voor hergebruik. In het verleden zijn grote open datasets gebruikt voor zowel commerciële gezichtsherkenningstoepassingen als voor toepassingen van politie en veiligheidsdiensten. We moeten een evenwicht zoeken tussen enerzijds het verzekeren van open data en transparante processen en anderzijds het voorkomen van ongewenst hergebruik.

BIAS

(Potentiële) vooringenomenheid van algoritmen noemen we bias. Binnen gezichtsherkenningstoepassingen komen vooral raciale en gender-biases voor. Bestaande gezichtsherkenningssystemen zijn vaak weinig transparant over de gebruikte algoritmen en baseren zich op trainingsdata

die onvoldoende representatief zijn. Zo zijn etnische minderheden, maar ook bijvoorbeeld vrouwen, kinderen en mensen met fysieke beperkingen vaak ondervertegenwoordigd.

Een zorgvuldige afweging tussen de potentiële winst die genderdetectie oplevert en de mogelijk kwalijke gevolgen ervan is cruciaal. Het algoritme kan immers fouten maken bij de genderbepaling, mede omdat niet iedereen zich herkent in de binaire opsplitsing tussen vrouw en man.

Door bias dreigt een bevestiging of uitvergroting van bestaande maatschappelijke ongelijkheden. Het is niet alleen van belang om, waar mogelijk, bias te voorkomen, maar ook om de gevolgen te proberen beperken en de potentiële bias maximaal zichtbaar te maken.

MILIEU-IMPACT

Gezichtsdetectie en -herkenning vragen veel rekenkracht en energie. Een substantieel deel daarvan komt niet uit hernieuwbare bronnen. Zowel het terugdringen van nefaste effecten (bijvoorbeeld door energie-efficiëntere workflows en algoritmen) als transparantie zijn belangrijke aandachtspunten.

DATA LABELERS

Algoritmen trainen met grote hoeveelheden data vraagt veel mankracht. Wie dat werk van data labeling uitvoert en in welke omstandigheden dat gebeurt, varieert. Soms gaat het om correct betaalde vaste werknemers, soms om stagiairs of vrijwilligers, al dan niet via crowdsourcing. Soms wordt ook een beroep gedaan op zogeheten *ghost workers* of *click workers* in lageloonlanden met penibele arbeidsomstandigheden.

Technologie is niet neutraal of onschuldig. Gezichtsherkenning verantwoord toepassen kan dus enkel met een ethisch kompas dat aandacht heeft voor de rechten en vrijheden van burgers. ■

Ethische aspecten

- **Automatisering als bedreiging voor jobs**

- focus: automatisering van werk dat vandaag wegens tijdgebrek / personeelstekort niet gebeurt;
- samenwerking tussen mens en algoritme blijft cruciaal.

- **Werkomstandigheden van *data labellers***

- we weten wie door ons gebruikte toolbox Insightface heeft ontwikkeld, maar niet met welk doel en in welke omstandigheden dat is gebeurd;
- aandachtspunt: manier waarop cultureelerfgoedorganisaties manuele validatie van matchingresultaten aanpakken.

- **Impact op het milieu**

- vergt veel rekenkracht en energie;
- energie mogelijk niet afkomstig uit hernieuwbare bronnen;
- aandachtspunten: terugdringen van nefaste effecten (bv. door energie-efficiëntere workflows en algoritmes) en transparantie.

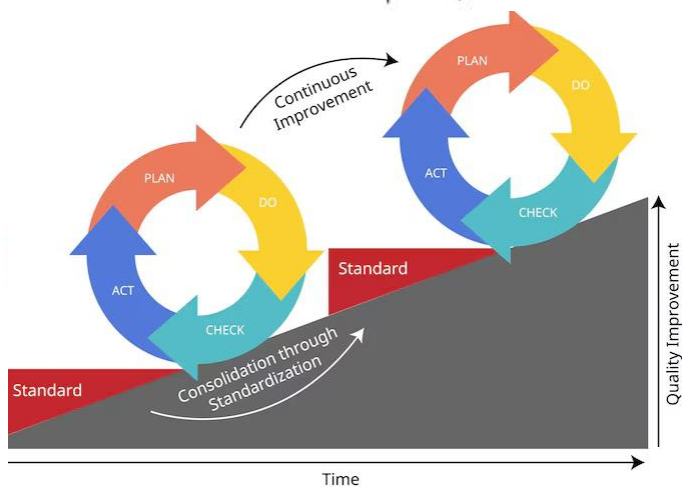
Lessons learned

- Bottlenecks:
 - samenstelling van referentiesets > tijdsintensief
 - manuele validatie > tijdsintensief
- Beschikbare informatiebeheersystemen zijn vandaag vaak niet voorzien op wegschrijven van dergelijke (complexe) beschrijvende metadata
- Juridische en ethische aspecten blijven belangrijke aandachtspunten

Ethische input voor volgend project

AANDACHTSPUNTEN

Principes/basishouding	Ethische doelen	Ethische knipperlichten
I. Prudentia (voorzichtig, verstandig, wijs) II. Iustitia (rechtvaardig, rechtschapen, juridisch kader, mensenrechten) III. Fortitudo (moed, doorzetting) IV. Temperantia (gematigd, proportioneel, zelfbeheersing) V. Plan - do - check - act-aanpak (FAME als eerste stap)	I. Negatieve impact voorkomen II. Negatieve gevolgen beperken (bv. validatie) III. Transparantie (over gebreken) en 'explainability' IV. Evenwicht belangen afgebeelde personen en belangen gebruikers/gemeenschap V. Rechtvaardigheid, objectiviteit, onpartijdigheid van archivaris VI. Het algemeen belang dienen VII. Gelijke toegang tot AI (voor publiek + voor collecties)	I. Bias (~ diversiteit, inclusiviteit) II. Privacy-impact op subjecten. Wat afgedekt door DPIA? Bv.: impact van combinatie van data, minderjarigen, niet-publieke personen, privacyvriendelijke softwarecomponenten,... III. Ethische werkomstandigheden (bv. data labellers) IV. Milieu-impact van tools en flows



GIVE metadata - scope

Spraakherkenning of **STT**

- Input: audio (uit audio/video)
- Output: transcripties

Entiteitsherkenning of **NER**

- Input: transcripties
- Output: metadata personen, locaties, organisaties incl. links naar authentieke bronnen (ef. wikidata)

Gezichtsherkenning

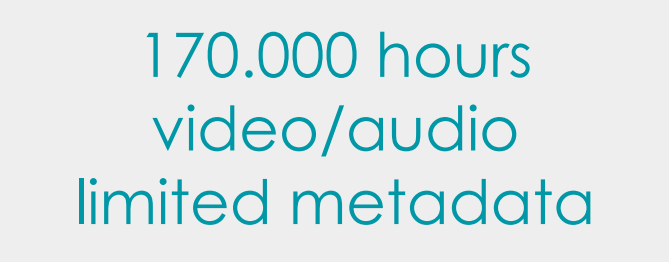
- Input: video
- Output: metadata herkende personen

Daarnaast

**Juridische
aspecten**

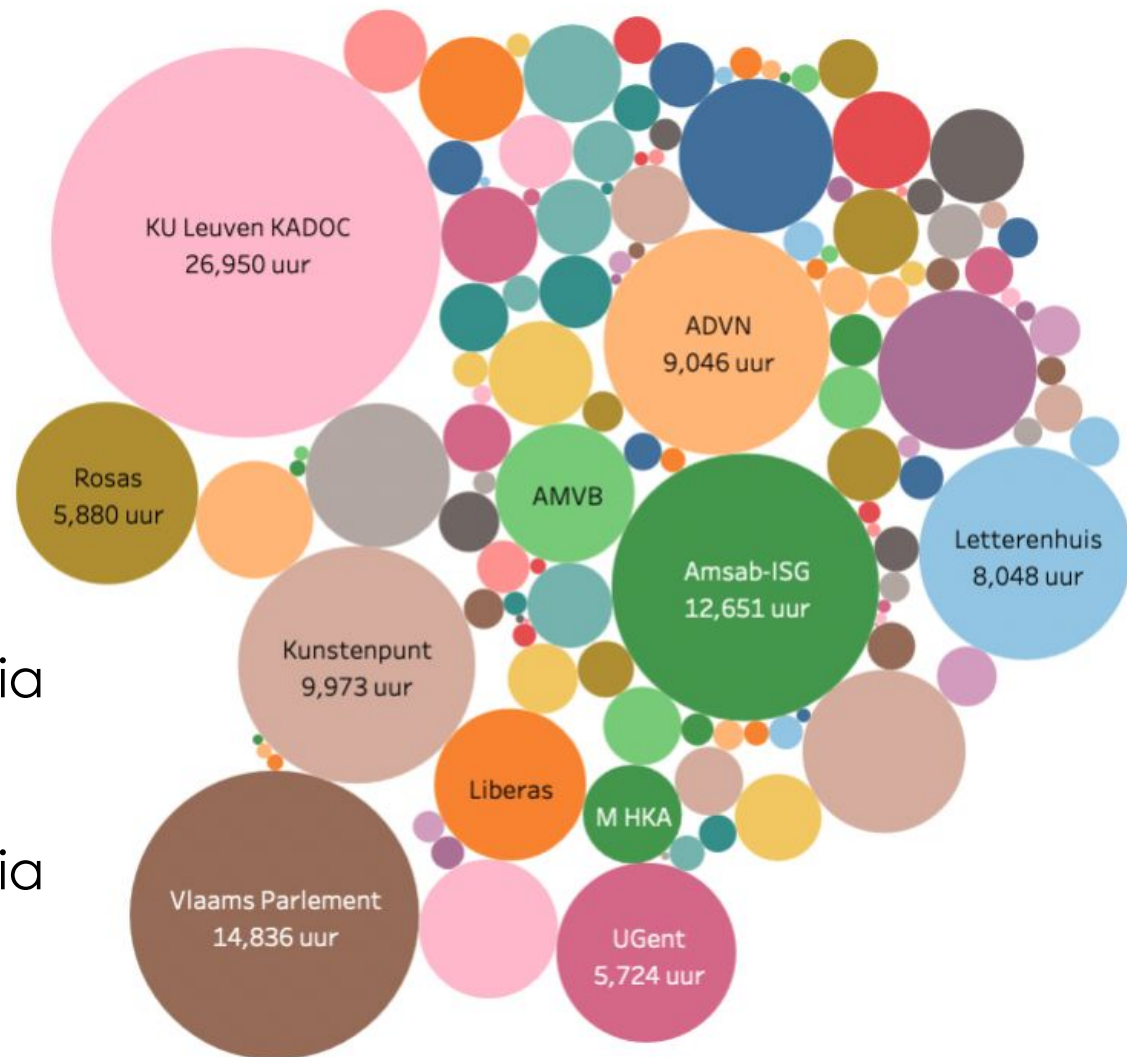
**Ethische
aspecten**

100



Werken op schaal

- 126 meemoo organisaties betrokken
- Cijfers:
 - STT en NER:
> 130.000 archief items, 170.000 uur media
 - Gezichtsherkenning :
> 100.000 archief items, 124.000 uur media



GIVE metadata - partner samenwerking

- We werkten met materiaal van heel veel partners
- Betrokkenheid via werkgroep
 - kregen meer in-depth informatie
 - Wat met nieuwe uitdagingen / inzichten
 - bepaalden mee wat we doen en niet doen
 - bv. beheer referentieset gezichten



GIVE metadata - Juridische aspecten

- Verder bouwend op juridisch kader FAME
 - Algemene Verordening Gegevensbescherming (AVG of **GDPR**)
 - Data Protection Impact Assessment (**DPIA**)
 - Omschrijven wat we wilden bereiken
 - Vooraf privacyrisico's van gegevensverwerking in kaart brengen
 - Maatregelen nemen om risico's te verkleinen
- > Beslissing om uit te voeren op basis van:
- Aantal betrokkenen
 - Volume van de gegevens
 - Duur van de activiteit

Data Protection Impact Assessment

- Deel 1 : omschrijf wat je wil doen

RISICO-ANALYSE							
	Verwerkingsactiviteit	Beschrijving risico		Schadetype	Toelichting type	Kans op schade	Ernst van de impact
	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	Beschrijf hier waarom je een bepaald type schade hebt aangeduid	Kies hier de kans dat de schade zich voordoet	Kies hier de impact van de schade
4				▼		▼	
5				▼		▼	
6				▼		▼	
7				▼		▼	
8				▼		▼	
9				▼		▼	

+ ≡

Gegevens DPIA ▼

Verwerking ▼

Risico-analyse ▼

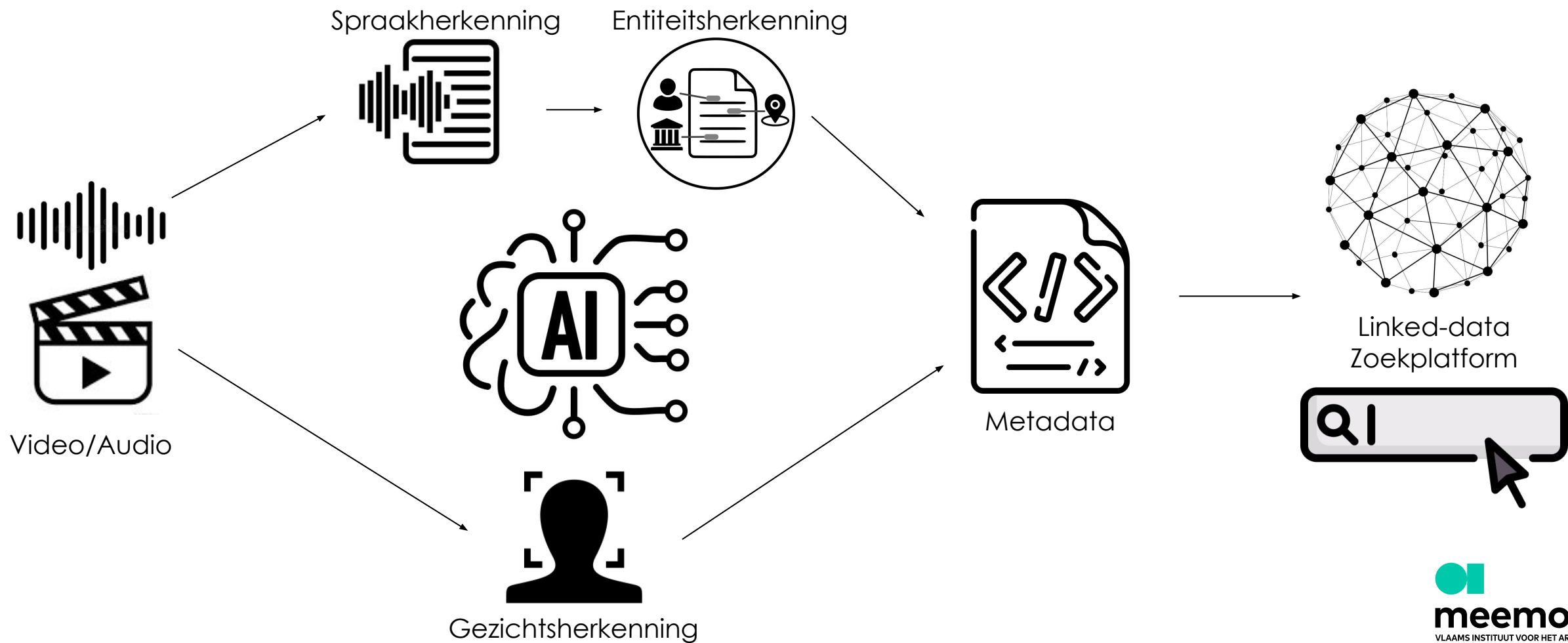
Detail Risico ▼

BIJLAGE 1 Voorbeelden GEGEVENS ▼

BIJLAGE 2 Voorbeelden GEGEVENS ▼

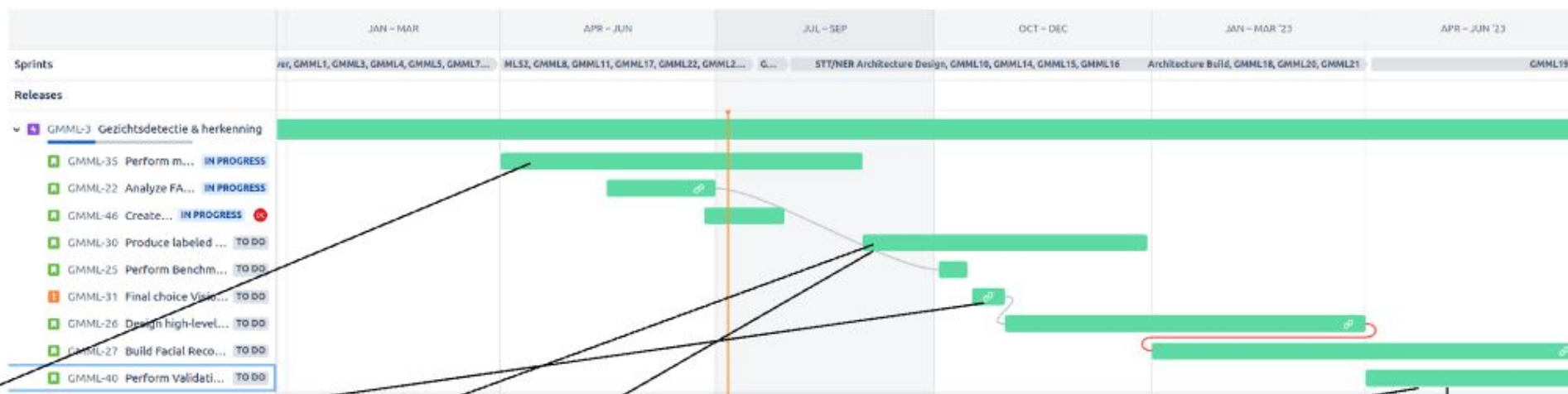
⇒ Checklist voor privacy aspecten binnen project

AI Metadata Pipeline



Ethische input

KADER ETHICAL IMPACT ASSESSMENT GIVE



Ethische 'agents'

- A. meemoo
- B. meemoo + CP's (+ stakeholders?)
- C. CP's

RFP
vendors
A

Keuze
model /
vendor
B

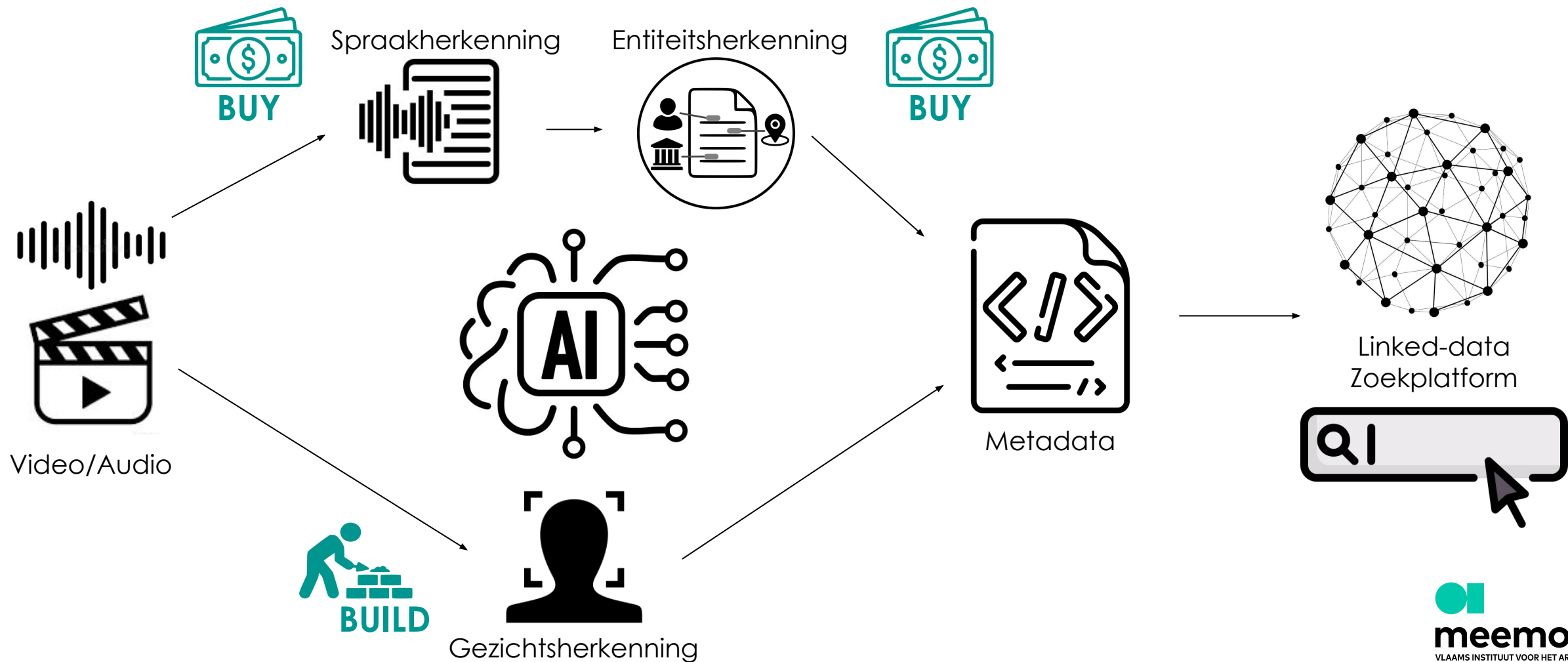
Lijst personen
samenstellen
B

Referentiesets
samenstellen
B

Algemene validatie
(performantie, bias,...)
A

Specifieke validatie
(thresholdbepaling,...)
C

AI Metadata Pipeline



Gezichtsherkenning - Build vs Buy

BUILD vs BUY

	Kosten	Effort	Technisch	Privacy
Build	+ Ruimte voor extra feature ontwikkeling	- Relatief grote investering	+ Keuze basic + extra features + Inzicht in de core algoritmes	+ Volledig in eigen handen
Buy	- Hoge prijs / uur - Post-GiVE kosten	+ SaaS oplossing klaar om te gebruiken - Mgmt platform dev kost	+ Continuous improvements SaaS updates - Features liggen ~ vast - Black box -> weinig ruimte voor tuning ?	- Niet in eigen handen

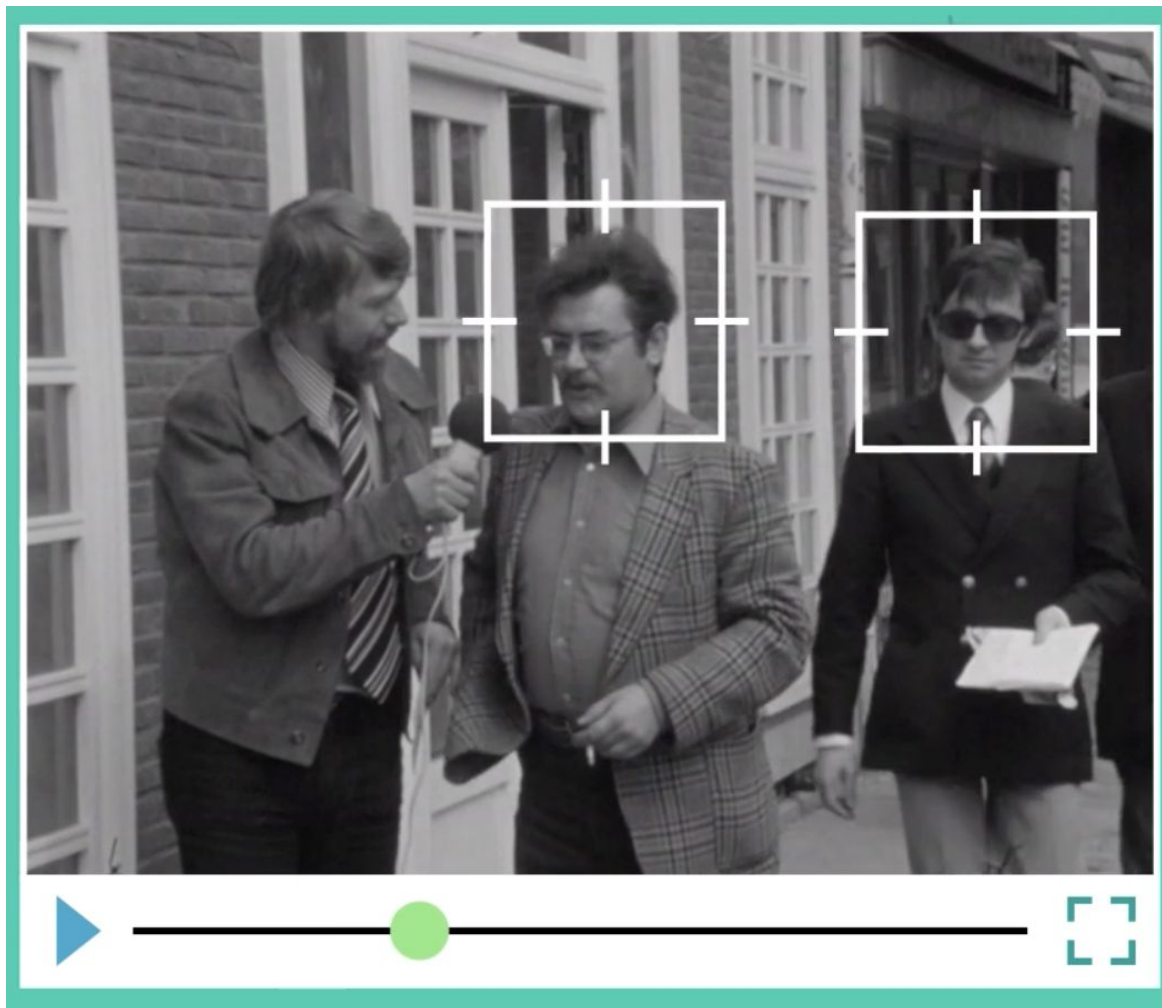
Conclusie: ontwikkelkost onder BUY kosten + ruimte extra onderzoek & features

→ we bouwen onze eigen oplossing

→ oplossing is herbruikbaar voor nieuw materiaal

(enkel compute- en onderhoudskosten)

Gezichtsherkenning - Wat willen we?



Gezichten detecteren en groeperen

Matchen met gezichten 'bekende' personen met **confidence score**

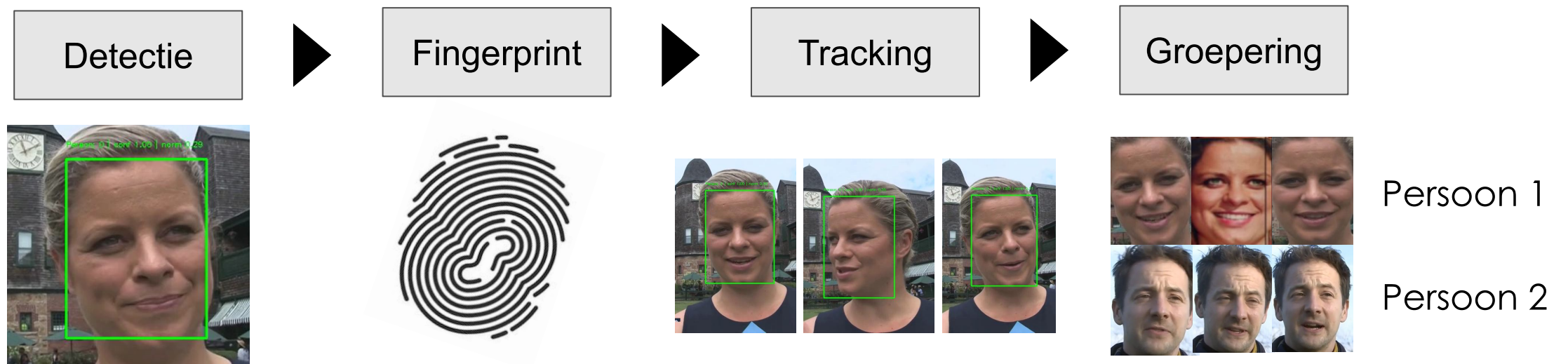
Identificeren wie vaak voorkomt maar niet gematcht kan worden

Beheer van een beperkte set gezichten (namen, foto's) die herkend mogen worden

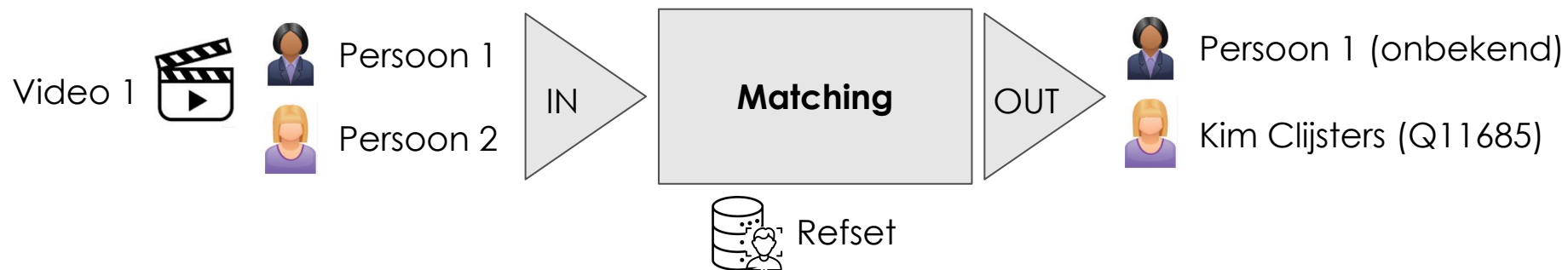
Gezichtsdetectie



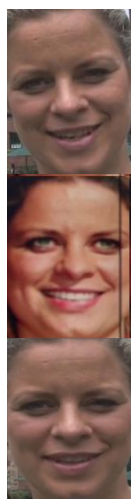
Detecteer gezichten in een video en **groepeer** de gezichten **per persoon**



Gezichtsherkenning (matching)



Match personen uit de video's met **bekende personen** in de **refset** → **Metadata**



Goede fingerprint gelijkenis



Kim Clijsters



Referentieset



Gezichten matchen met de referentieset

AMSAB
Archief



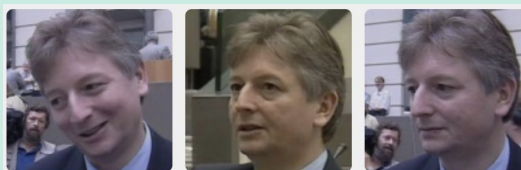
Output Gezichtsdetectie



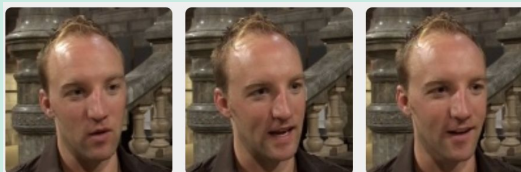
Tijd in beeld: 4m 35s



Tijd in beeld: 0m 8s



Tijd in beeld: 0m 32s



Tijd in beeld: 0m 39s



Matching



Onbekenden
Clustering



Gedeelde Referentie Set



Steve Stevaert

Politicus

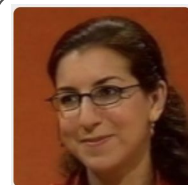
Wikidata: Q1234



Joke Devynck

Actrice

Wikidata: Q5678



Fauyaza Talhaoui

Politica

Wikidata: Q1011

Onbekende gezichten! Bel een archivaris

AMSAB
Archief



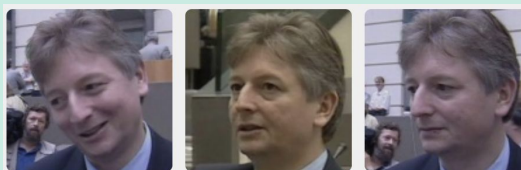
Output Gezichtsdetectie



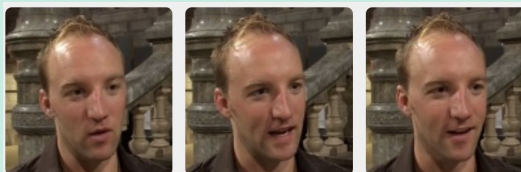
Tijd in beeld: 4m 35s



Tijd in beeld: 0m 8s



Tijd in beeld: 0m 32s



Tijd in beeld: 0m 39s

Gedeelde Referentie Set



Steve Stevaert

Politicus

Wikidata: Q1234



Joke Devynck

Actrice

Wikidata: Q5678



Fauyaza Talhaoui

Politica

Wikidata: Q1011



Mathijs Scheepers

Acteur

Wikidata: Q3389

Onbekenden
Clustering



Gedeeld beheer van de referentie set

Wat is dit?

- Lijst van personen + links naar Wikidata + foto's
- Gedeeld beheer, gebruikmaken van expertise in sectoren
- Transparantie is belangrijk (wie deed wat wanneer?)

Cruciaal element in onze processen

- Brug tussen detectie en matching
- **Ethisch aspect: iemand toevoegen aan de referentieset betekent deze op schaal identificeren. Hierover moet een mens beslissen.**

AI tool positionering



- aparte tooling
- experimenteren: niet automatisch publiek doorzoekbare metadata
- evaluatie van kwaliteit
- rol van collectiebeheerders/archivarissen: beslissen welke metadata wordt opgenomen

Gezichtsherkenning - Bias analyse

Mogelijke bronnen van bias

- Gezichtsdetectie: lagere detectierate
- Gezichtsherkenning: lagere herkenningsrate

Onderzoek (op vraag van werkgroep)

- Moeilijk om harde cijfers te verkrijgen
 - Nood aan gelabelde dataset (hoge kost)
 - Evaluatie is complex

→ Praktische aanpak: evaluatie adhv resultaten



Claron McFadden (Q454090)

Tan Dun (Q314176)

Souad Boukhatem (Q19971918)

Conclusie

We zijn ervan overtuigd dat de **gezichtsherkenning goed werkt** in **uitdagende situaties** voor **maximale diversiteit**. Het blijft echter **moeilijk** om dit **te bewijzen** met hard **cijfermateriaal**.

What's cooking?



Mona Hassan Abo-Abda, [CC BY-SA 4.0](#), via Wikimedia Commons

Gerelateerde projecten

- **SHARED-AI & HERMES (opvolgprojecten)**
 - Meer partners (mediasector bijvoorbeeld)
 - Meer technieken (stemherkenning, correcties op metadata, ...)
- **Visual Name Authority**
 - 2023 - 2026 (3 fases, 3 jaar)
 - Project waarin we overkoepelende namenlijst en gedeelde referentieset breder inzetbaar willen maken
 - Processen, juridisch, ethische kwesties
 - bouw/aankoop van tool voor beheer

Beleidsplan 2026-2030: van project naar duurzame werking?



Bart Magnus
bart.magnus@meemoo.be

Bedankt! Vragen?