Verslag Proof of Concept Personenthesaurus

Datum: oktober 2024
Auteurs: Eric van Balkum, Thomas Op de Coul, Kathrin Dentler, Nynke Kuipers, Ruben Schalk, Elena Slavco, Mirjam Verloop en Mari Wigham

podium kunst .net

Inhoudsopgave

1. Inleiding

1.1. Achtergrond

Terminologiebronnen spelen een belangrijke rol bij het verbinden van collecties, in het bijzonder de bronnen met personen (in vrijwel alle collecties komen tenslotte personen voor). Er bestaan al verschillende toegankelijke bronnen met informatie over personen. Daarnaast hebben sommige organisaties een rijke, maar nog niet gepubliceerde bron. Ook zijn er organisaties die nieuwe bronnen ontwikkelen, bijvoorbeeld vanuit een geografisch of domeinspecifiek oogpunt. Hoe brengen we deze bronnen samen en publiceren we die als één bron met personen uit de podiumkunsten?

In recente bijeenkomsten van de Werkgroep Terminologiebronnen is nagedacht over een aanpak. Het voorstel van de werkgroep is om een virtuele bron of virtuele thesaurus te maken. In die opzet blijven alle bronhouders verantwoordelijk voor hun gegevens, en dient de virtuele bron als 'verzamelpunt' voor alle verbindingen tussen de verschillende bronnen.

Bijvoorbeeld, een terminologiebron van Muziekweb bevat personen en een terminologiebron van Muziekschatten bevat ook personen. De virtuele bron verwijst naar en verbindt dezelfde personen uit beide terminologiebronnen. Een eindgebruiker, bijvoorbeeld een archivaris of metadateerder, gebruikt enkel de virtuele bron.

Uitgangspunten:

1. De centrale thesaurus voor personen verbindt alle beschikbare informatie over personen uit de podiumkunsten.
2. De centrale thesaurus voor personen maakt het voor erfgoedinstellingen makkelijker om verwijzingen naar personen toe te voegen aan hun collecties.
3. De centrale thesaurus voor personen helpt bij het verbinden van erfgoedcollecties van verschillende podiumkunstinstellingen.
4. De centrale thesaurus voor personen is bruikbaar voor het brede erfgoedveld en beschikbaar via het Termennetwerk.
5. De aanpak en de gemaakte (technische) keuzes zijn herbruikbaar voor andere thesauri, zoals bijvoorbeeld een thesaurus met werken.
6. Bestaande bronnen vormen de basis voor de centrale thesaurus voor personen. Er is een werkwijze voor het toevoegen van nieuwe bronnen of nieuwe termen.

1.2. Doel

- Welke benadering werkt efficiënt om personenbronnen 'uniform' te gebruiken?
- Technisch creëren van een centrale thesaurus (binnen de scope van een *Proof of Concept*: een showcase van hoe het werkt)

1.3. Scope

Het doel van dit project is nadrukkelijk niet om te komen tot een bruikbare of een 'definitieve' variant van een centrale thesaurus voor personen. Dit project is een *Proof of Concept*. Dat betekent dat de nadruk ligt op het selecteren van werkwijze en hulpmiddelen, en het beschrijven en toetsen van een aanpak. Het projectteam werkt met twee of drie vooraf gekozen terminologiebronnen.

1.4. Projectteam

De uitvoering, visie en kwaliteitscontrole is vanuit het projectteam neergezet.
Het projectteam kwam gedurende 6 weken tweemaal per week digitaal bijeen. Elena en Kathrin hebben het uitvoerende werk gedaan. De resultaten werden besproken en inhoudelijk getoetst.

Door de korte looptijd en actieve benadering hebben we in korte tijd veel resultaat weten te behalen. Deze aanpak wordt door alle betrokkenen als erg prettig ervaren. Er kwamen goede ideeën, discussies en concrete vervolgstappen uit elke sessie.

Samenstelling projectteam:

- Nynke Kuipers (product owner namens podiumkunst.net)
- Mirjam Verloop (podiumkunst.net)
- Eric van Balkum (muziekschatten.nl/podiumkunst.net)
- Thomas Op de Coul (Beeld & Geluid, muziekweb.nl)
- Mari Wigham (Beeld & Geluid, GTAA)
- Ruben Schalk (RCE, Termennetwerk)
- Elena Slavco (Triply)
- Kathrin Dentler (Triply)

2. Construeren Personenthesaurus

2.1. Links

- *De dataset om de thesaurus te construeren: https://podiumkunst.triply.cc/Personenthesaurus/Construct-Thesaurus*
- *De thesaurus zelf: https://podiumkunst.triply.cc/Personenthesaurus/Thesaurus*
- *Data story: https://podiumkunst.triply.cc/Personenthesaurus/-/stories/Personenthesaurus-Podiumkunst*

2.2. Aanpak

De dataset Construct-Thesaurus bevat de relevante data van de drie bronnen: Muziekschatten, Muziekweb en GTAA. Om time-outs tijdens federatieve bevragingen te voorkomen, worden de nodige triples in scripts op basis van gepagineerde SPARQL queries geëxtraheerd en vervolgens naar de dataset geüpload. Vervolgens wordt de thesaurus puur op basis van SPARQL queries opgebouwd.

Eerste stap: datakwaliteit en verrijkingen.

- Een query selecteert alle Wikidata links in Muziekweb en voegt nieuwe links toe die zoals de andere bronnen het prefix van de Wikidata entiteiten gebruiken (Zowel de http:// als ook de https:// versie, dus https://www.wikidata.org/entity/ en http://www.wikidata.org/entity/, Muziekweb zelf gebruikt het prefix https://www.wikidata.org/wiki/).
- Een tweede query normaliseert alle persoonsnamen. Hiervoor worden de namen op basis van de eigenschappen skos:prefLabel, skos:altLabel, sdo:alternateName en sdo:name geselecteerd. Vervolgens worden de volgende stappen doorlopen:
    - Namen worden van “Achternaam, Voornaam” naar het patroon “Voornaam Achternaam” genormaliseerd.
    - Namen worden geconverteerd in strings (dus zonder language tags) en kleine letters, en diakrieten worden door letters zonder diakrieten vervangen.
    - (We hebben met de optie geëxperimenteerd initialen te verwijderen, maar dat had maar een beperkt effect en resulteerde in een lagere kwaliteit).
    - Alle tekens die geen letters zijn worden verwijderd.
    - Tenslotte worden lege genormaliseerde namen eruit gefilterd.
    - Om de genormaliseerde namen van brondata te kunnen onderscheiden worden ze met de eigenschap schema:callSign in de dataset opgenomen.

Tweede stap: voor-berekening van relaties.

- Een query voegt relaties aan de dataset toe.
    - Directe relaties, als een bron expliciet met een owl:sameAs of skos:exactMatch naar een persoon in een andere bron verwijst.
    - Indirecte relaties, als twee bronnen allebei met een owl:sameAs of skos:exactMatch naar dezelfde persoonsentiteit in bijvoorbeeld Wikidata of Discogs verwijzen, of naar hetzelfde callSign[1].
    - Tenslotte wordt voor iedere persoon ook een relatie met zichzelf toegevoegd. Om technische redenen is dit een noodzakelijke stap om de volgende query goed te laten werken.
    - Om de toegevoegde relaties van relaties in de brondata te kunnen onderscheiden, worden ze met de eigenschap schema:relatedLink in de dataset opgenomen.

Derde stap. De Thesaurus wordt geconstrueerd.

- Een query voor alle “clusters”. Een cluster is een groep van personen die met schema:relatedLink eigenschappen aan elkaar verbonden zijn, en waarbij geen tegenstrijdige geboortejaren in de data zitten. We beschouwen alle personen in een cluster als dezelfde persoon. De query wordt twee keer gedraaid: een keer naar de dataset Thesaurus en dan naar Construct-Thesaurus.
- Een query die instanties voor de overige personen in Muziekschatten en Muziekweb construeert[2]. Deze personen konden niet geclusterd worden. Het kan zijn dat deze uniek zijn binnen de groep van bronnen. Het kan ook zijn dat sommige van deze personen wel degelijk ‘dezelfde persoon’ zijn als een cluster of een andere overige persoon, maar dat dit niet vastgesteld kon worden op basis van onze criteria. De resultaten van deze query gaan eerst naar de dataset Thesaurus en dan naar Construct-Thesaurus.
- En de laatste query voor verrijkingen, om bij de personen in de thesaurus een visueel overzicht te geven uit welke bronnen de *triples* komen. De resultaten van deze query gaan direct naar de dataset Thesaurus.

[1] Alleen voor Muziekweb en Muziekschatten. Doordat de GTAA meerdere domeinen bevat levert matching op hetzelfde *callSign* alleen te veel verkeerde matches op.
[2] Niet voor de GTAA omdat de scope van deze thesaurus veel breder is dan het muziekdomein alleen, waardoor er te veel ruis ontstaat.

3. Lessons learned

3.1. Het is mogelijk een Personenthesaurus te construeren

Het is gelukt om een 'virtuele personenthesaurus' te realiseren die samengesteld is uit meerdere bronnen. In deze PoC is gekozen voor een domeinspecifieke aanpak door de focus te leggen op het domein 'muziek' binnen de podiumkunsten.

Er zijn in deze PoC drie verschillende Linked Data bronnen gebruikt: Muziekweb, Muziekschatten en GTAA personen (Beeld en Geluid).

Het is gelukt om de 'virtuele Personenthesaurus' op te bouwen door middel van een geautomatiseerd proces van SPARQL-queries. Ook om de namen te normaliseren (vergelijken zonder afleiding van leestekens) blijkt de toegepaste SPARQL-query erg effectief.

Een van de uitgangspunten van de werkgroep is om geen redactie op het samenstellingsproces te hoeven voeren. Door verschillende kwaliteitscontroles tijdens het proces van de PoC zijn de SPARQL-queries zodanig aangescherpt dat de uitkomst van het geautomatiseerde proces een hoge mate van betrouwbaarheid heeft.

3.2. Conceptueel omgaan met grote hoeveelheid bronnen en data

Conceptueel is het een ingewikkelde uitdaging om met zoveel bronnen en data om te gaan. Hier is dan ook veel tijd in gaan zitten om tot een acceptabel concept te komen.

Het oorspronkelijke plan was om de records paarsgewijs (per tweetallen) te vergelijken. Dat werkt alleen niet goed als de matching op basis van 'geen tegenspraken' werkt. Stel dat een persoon zonder geboortejaar eerst met een matching kandidaat mét een geboortejaar wordt vergeleken en dan met een andere matching kandidaat die een ander geboortejaar heeft. Dan zouden we twee paarsgewijze matches hebben, maar een tegenspraak in het cluster. Dit is opgelost door de eis dat de verbindende persoon in een cluster altijd een geboortejaar moet hebben. Een strengere manier waarin we altijd een geboortejaar eisen zou minder complex zijn.

De verwachting was dat SPARQL niet toereikend zou zijn. De PoC Personenthesaurus laat het tegendeel zien. Door tussenberekeningen in Linked Data op te slaan is het mogelijk om de hele Personenthesaurus door middel van SPARQL samen te stellen. Er zijn 9 SPARQL-queries om de Personenthesaurus op te bouwen (3 voor de import van data van de drie bronnen, 2 om data te verrijken, 1 voor de relaties, en 3 om de Personenthesaurus op te bouwen). Tijdens de PoC zijn we niet tegen time-outs aangelopen. Op dit moment is het draaien van de queries nog handwerk, de wens is om dit om te zetten naar een geautomatiseerd proces.

Deze aanpak komt naar ons idee de implementeerbaarheid en bruikbaarheid van de tool ten goede. We hebben gebruik kunnen maken van een gangbare techniek om Linked Data te verwerken. Hierdoor hoeven we niet te grijpen naar ingewikkelde (en specifiek voor dit doel gemaakte) scripts waar zeer specifieke programmeerkennis voor nodig is om dit in te kunnen zetten.

3.3. Kwaliteit is belangrijker dan kwantiteit

Om als 'virtuele Personenthesaurus' een meerwaarde te creëren is kwaliteit belangrijker dan kwantiteit. Immers, de verschillende bronnen zijn op zichzelf al beschikbaar. Door de bronnen samen te brengen en dezelfde personen in 'clusters' te presenteren wordt het gebruiksgemak verbeterd. Om dit te realiseren is het wel noodzakelijk om hoog-kwalitatieve matches te realiseren.

Bij kwantiteit breng je de bronnen samen, maar ontstaat er heel veel ruis door personen die niet of niet betrouwbaar gematcht worden. Zo ontstaan mogelijk veel dubbele persoons-entiteiten die eigenlijk over dezelfde persoon gaan. Of er ontstaan juist heel onbetrouwbare clusters met personen die niets met elkaar te maken hebben, behalve de naam die ze dragen.

Om tot een kwalitatief hoogwaardige 'virtuele Personenthesaurus' te komen is het noodzakelijk om eisen aan de brondata te stellen. Zo dient de bron in Linked Data beschikbaar te zijn en moeten er minimale velden worden meegegeven. De PoC heeft ons laten zien dat rijke, gecureerde bronnen vele malen betrouwbaarder te matchen zijn dan meer algemene bronnen.

3.4. Naam alleen is niet genoeg

Personen louter matchen op genormaliseerde namen levert heel veel onbetrouwbare matches op. Bij de aanpak is aanvullende informatie naast de naam een *must have*.

In de PoC hebben we de strategie gebruikt om te matchen op relaties en genormaliseerde namen met een controle op geboortejaar (marge van 3 jaar). Dit levert dan wel niet de meeste matches op, maar wel de meest betrouwbare.

Hierdoor is het de aanbeveling vanuit de PoC om de volgende velden als minimale eis te stellen aan terminologiebronnen (eis wordt om aan één van deze velden te voldoen):

- IRI naar een externe bron. De verwijzing naar een externe bron kan je heel goed opnemen in je eigen database. Bij deze PoC zijn hiervoor de velden owl:sameAs en skos:exactMatch binnengehaald bij de bronnen.

- Geboortejaar (mag ook een exacte datum zijn). Deze informatie zorgt ervoor dat er op twee feiten gematcht kan worden: naam en jaar. Daarnaast is het ook voor intern gebruik erg waardevol, denk bijvoorbeeld aan het bepalen van de auteursrechten.

“Je hebt vaak terminologiebronnen nodig om een terminologiebron te maken” (Eric van Balkum)

Wat we in de PoC heel erg hebben gemist is het kunnen inzetten van een veld ‘rol’, simpelweg omdat deze in de verschillende bronnen geen eenduidige informatie bevatte. Zo heeft de GTAA wel een scope note, maar die is niet gestandaardiseerd ingevuld. Hierdoor was het bij de GTAA niet mogelijk om specifiek te filteren op alleen de personen binnen het domein ‘muziek’. Muziekweb en Muziekschatten samen blijken erg betrouwbaar te matchen omdat deze beide een focus hebben op hetzelfde domein.

Dit vraagstuk kan ook worden opgelost door vanuit de hiërarchie van een bronthesaurus te kunnen herleiden welk deel je wilt gebruiken in de thesaurus. Dit is bijvoorbeeld in het Termennetwerk gebruikt om van de AAT de ‘materialen’ apart te bevragen.

Een derde optie is om subsets van een thesaurus te gebruiken, bepaald op basis van aan welk materiaal ze worden verbonden in het archief. Bijv. alle personen uit de GTAA die aan programma's met genre 'muziek' worden verbonden.

3.5. Zoveel bronnen, zoveel verschillen

Zelfs met Linked Data blijkt het nog niet zo eenvoudig om meerdere bronnen samen te brengen. Maar omdat we de Linked Data techniek gebruiken in het samenstellen van de ‘virtuele Personenthesaurus’, detecteren we eventuele inconsistenties. Door bij de bron deze data (in de Linked Data) te verbeteren of dit in de SPARQL-query consequent op te lossen wordt het proces steeds eenvoudiger.

Wat voorbeelden die we zijn tegengekomen:

- Verschillen in het gebruik van predicaten voor dezelfde informatie (bijv. owl:sameAs en skos:exactMatch)
- Verschillende informatie in één veld. (Bijv. Muziekweb maakt in rdfs:type geen onderscheid tussen naam (persoon) en naam (band), alle namen hebben zowel schema:person als schema:musicGroup)
- Uniform gebruik van velden: gebruik voor het datumveld alleen ISO standaard voor datum (Jaar altijd 3 of 4 getallen; maand en dag altijd 2 getallen vb: YYYY-MM-DD). Gebruik bij voorkeur geen letters of leestekens. Deze notaties worden niet meegenomen in het matchen.
- Verschillen in IRI notaties (bijvoorbeeld https://www.wikidata.org/wiki/ en http://www.wikidata.org/entity/) of http:// en https://).

- Alleen specifieke subset van een bron mee kunnen nemen
    - Bij een algemene bron kan het bijvoorbeeld gaan om een selectie van een specifiek domein (bv. bij GTAA niet echt mogelijk)
    - Voorbeeld Muziekschatten: Filter “som:ZKNMFZ” is ingevuld. Het gaat hier om het veld dat de ‘rol’ van de persoon aangeeft. Als dit veld gevuld is, dan wel meenemen. Is het veld niet gevuld: dan niet meenemen in data (want niet-gecureerd).
    - Let op: die subset moet in SPARQL bevraagd kunnen worden, dus geen aparte dump.

3.6. Foutdetectie op brondata

Een heel mooie toepassing is dat de matching helpt om kwaliteitsissues bij de bron op te sporen. Als twee concepten via een directe of indirecte link van hoge betrouwbaarheid (bijvoorbeeld owl:sameAs) met elkaar gematcht zijn maar de geboortedata niet kloppen dan is de kans groot dat in de brondata een fout is gemaakt. Deze resultaten kunnen worden teruggegeven aan de bronhouder.

3.7. Zoek Vader Abraham en vind Pierre Kartner

Hoewel we de pseudoniemen buiten de scope van de PoC hebben gehouden, is er al een goed resultaat te melden vanuit de uitgewerkte benadering. Ook als je zoekt op een pseudoniem komt je uit op één persoon - de ‘hoofdpersoon’.

Via de bronhouder kan je achterhalen welke alternatieve naam/pseudoniem verwijst naar werk dat onder dat pseudoniem is gemaakt. Een aantal voorbeelden van pseudoniemen:

- Zangeres zonder naam
- Vader Abraham
- Jan Stoeckart
- Jacques van Tol
- Drs. P

3.8. Neverending story

Het proces van optimaliseren van de matching van personen en het construeren van een Personenthesaurus is een *neverending story*. Zo biedt iedere oplossing weer nieuwe kansen voor verdere *finetuning* van de ‘virtuele Personenthesaurus’. Verbeteringen in brondata bieden ook weer nieuwe kansen voor *finetuning*. Denk aan opgeloste ‘fouten’, toegevoegde IRI’s en geboortejaren.

Het is een afweging tussen kosten en baten, die altijd geïnformeerd moet worden door de ervaring van de gebruiker, hoe ver je hiermee wilt gaan.

4. Vragen en acties voor een vervolg

- Voorziet deze Personenthesaurus in de behoefte die er is?
- Voor welke doelgroep(en) wordt de Personenthesaurus ingezet?
    - Eindgebruiker (collectiebeheerder/registrator)
    - Ontwikkelaars (de gegevens overnemen en integreren in eigen toepassing)
- Is de huidige manier waarop IRI's worden gegenereerd gewenst? NDE-proof?
- Hoe gaan we de thesaurus gebruiken?
    - Geconstrueerde digitale Personenthesaurus (zoals gerealiseerd in de PoC)
    - Of wordt de Personenthesaurus meer gebruikt als entrypoint op de verschillende bronnen?
    - Of Personenthesaurus als tussencachingslaag. Zoeken op een string -> wij suggereren dat daar deze persoon/personen uit komt met deze IRI's vanuit de verschillende bronnen. Keuze aan de gebruiker welke IRI gebruikt wordt. (Termennetwerk +)
    - Hoe is de implementeerbaarheid van een mogelijke benadering? De benadering die de meeste resultaten oplevert is niet per se de benadering die eenvoudig te implementeren is voor het erfgoedveld. Vooral de implementatie is relevant voor deze POC.
- Hoe gaan we de thesaurus bevragen?
    - Via een API of handmatig?
    - Elasticsearch werkt goed, maar we kunnen de ranking niet beïnvloeden
    - Het zou een optie zijn om met SPARQL te zoeken, en van de input ook een "callSign" te maken.
- Vervuiling, onvolledigheden of onjuistheden kunnen worden geconstateerd in het proces en wellicht teruggegeven worden aan de bronhouder
    - Sommige personen zijn geen persoon
    - Bijvoorbeeld twee namen die genormaliseerd gelijk zijn, maar twee bron IRI's hebben (vb GTAA: Fréquin, Willibrord en Frequin, Willibrord)
    - Vervuiling van de bron komt natuurlijk ook in de virtuele thesaurus terecht > Vervuiling dient altijd bij de bron opgelost te worden, niet in de virtuele thesaurus
- Hoe ver gaan we met het verrijken van externe bronnen?
    - Is er een geautomatiseerde manier te ontwikkelen die de geboortedata toevoegt aan de personen en er zo gewerkt kan worden met verrijkte data voor het matchen? En is deze geautomatiseerde manier dan goed genoeg dat we de geboortedata voldoende vertrouwen?
    - De meerwaarde van dit proces zit in de kwaliteit van de match én de kwaliteit van de externe bron.
    - Is dit wenselijk voor bronhouders? Of gaat dit in tegen afspraken bij de bronhouder? Denk aan vraagstukken over dupliceren van informatie en sync issues.

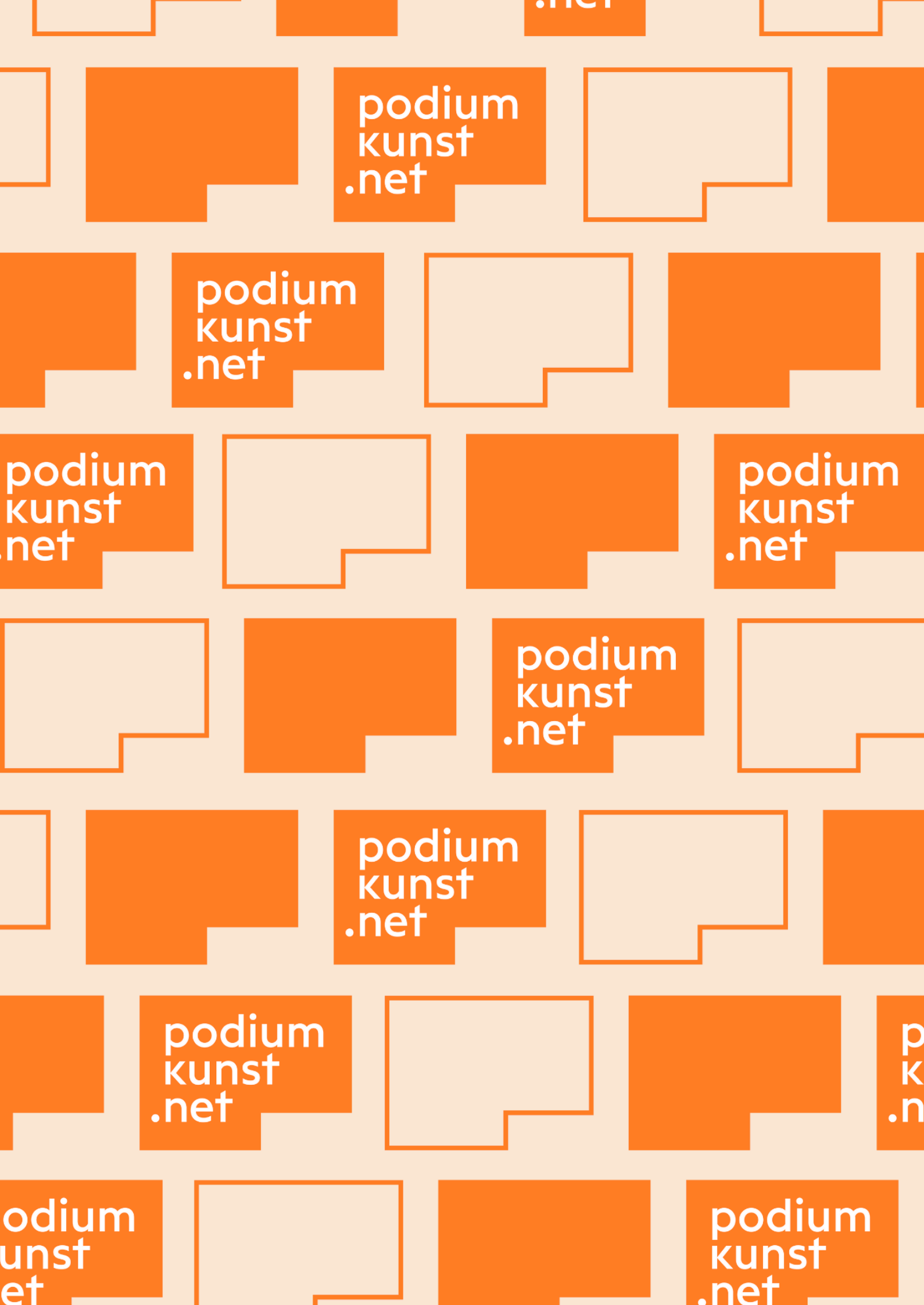
    - Hoe kan dit proces betrouwbaar uitgevoerd worden?
    - Hoe ver willen we gaan? Dit is echt nog een heel open discussie met veel haken en ogen.
- Wie gaat het beheer van de 'Virtuele Thesaurus' voeren?
    - Wie gaat het beheer voeren over de te realiseren virtuele thesaurus?
    - Welke implicaties heeft dit voor de organisatie die dit op zich gaat nemen?
    - Wat betekent het voor het beheer?
    - Wat betekent het voor kosten?
- Wat valt inhoudelijk onder het beheer van de Personenthesaurus?
    - Nieuwe bronnen toevoegen
    - Query's om brondata op te halen, te matchen en te construeren is op dit moment nog een handmatige actie
    - Aanpassingen / edits in data die invloed hebben op de Personenthesaurus:
        - Verwijderde of bijgewerkte personen uit bronnen (bron IRI's)
        - Foutieve samenvoegingen
        - Losse entiteiten worden later toch samengevoegd
        - Hoe behoud je de IRI's?
            - Bij het Termennetwerk is er een Lookup functie die checkt of de gebruikte URI's nog werken.
            - Bij ISNI handmatige aanpassingen mogelijk. Automatische checks waarbij e-mail verstuurd wordt over 'deze lijken op elkaar' etc.
    - Redactie / kwaliteitscontroles. De werkwijze moet duidelijk gecommuniceerd worden, zodat het voor gebruikers duidelijk is wat ze kunnen verwachten.
- Fuzzy matching
    - In de PoC is een test gedaan. Hier kwamen in een (eerdere) fase 84 matches op de fuzzy match uit ten opzichte van ca. 10.000 matches via de IRI en de naam exact.
    - Gezien de stap van Fuzzy Matching (Levenshteinafstand) een tijdrovende stap is met veel handwerk en scripting, is gezien de getallen gekozen om dit verder in de PoC links te laten liggen.
    - Hoewel voor deze POC dit veel te weinig matches oplevert, kan deze derde stap zeker te adviseren zijn als in de eerste twee stappen te weinig matches worden gevonden en als de fuzzy matching voldoende kwaliteit oplevert.
- Is het realistisch om bronnen 'eisen' op te leggen
    - Need to have: naam, geboortejaar en externe bron (of 'rol'?)
    - Bijvoorbeeld in het geval van de GTAA gaat het toevoegen van geboortejaren in tegen de strategie van Beeld & Geluid.
    - Hoe om te gaan met bronnen die niet volledig aan de eisen (kunnen of willen) voldoen?
    - Geboortejaar meenemen > er worden bronnen uitgesloten, maar matches zijn zekerder

    - Geboortejaar niet meenemen > bronnen kunnen altijd meegenomen worden, maar matches zijn onduidelijker
    - Testen in twee groepen: welke situatie wordt als 'irritanter' gezien?
- Hoe domeinspecifieke personen bijeen te brengen?
    - De bronnen die gebruikt kunnen worden in een 'virtuele Personenthesaurus' zijn lang niet altijd toegespitst op het domein van de 'podiumkunsten'. Het werken vanuit een domeinspecifieke thesaurus werkt goed omdat de gebruiker meteen weet dat het alleen om 'podiumkunsten' gaat. Daarnaast kan er betrouwbaarder gematcht worden als bijvoorbeeld 'biologen' met mogelijk dezelfde naam uitgesloten kunnen worden.
    - Het veld 'rol' zou daar een heel belangrijke rol in kunnen spelen. Tijdens de PoC is dit een veld waar inhoudelijk meteen naar gekeken en teruggegrepen wordt, echter dit is een handmatige evaluatie.
    - Op dit moment zit deze informatie op verschillende plekken in de brondata (bv GTAA in scopeNote als 'string' die heel vrij wordt ingevuld)
    - Vereiste is dan wel dat dit uniform ingevoerde data is (dus vanuit vooraf gedefinieerde lijsten).
    - Optie met de Theaterencyclopedie?
    - Er kan nog een test gedaan worden met een lijst met gewenste rollen en die matchen aan brondata (zoals scopeNote) om zonder aanpassingen in de bron toch de informatie uniform in te kunnen zetten
- Welke (andere) velden wil/kan je nog meenemen?
    - hiddenLabel > wordt nu nog niets mee gedaan
    - Publieksvelden: welke informatie wil je laten zien aan de gebruiker van de thesaurus? (Geboorteplaats?/Plaats van overlijden?/Nationaliteit? dan nieuwe terminologiebronnen vaststellen voor deze gegevens t.b.v. IRIs (geen *literals*))
    - Percentage van de zekerheid van een match?
- Hoe gaan we om met pseudoniemen of fictieve personen?
    - Ondanks de positieve 'bijvangst' moet er nog uitgebreid ingezoomd worden op de vraag hoe om te gaan met pseudoniemen
    - Alternatieve benamingen in Muziekschatten - 1 veld, 2 benaderingen
        - Een pseudoniem krijgt een eigen entiteit, binnen de database wordt in 'alternatieve naam' verwezen naar ander record
        - Een afwijkende naamvorm krijgt geen eigen entiteit, maar staat geschreven in 'alternatieve naam'
    - Eerst skos:prefLabel daarna pas skos:altLabel
    - skos:prefLabel matches betrouwbaar, skos:altLabel pas toevoegen na controle
- De VIAF vs ISNI discussie: tot nu toe hebben we nog geen stappen ondernomen om de meest aannemelijke waarden voor een nader te bepalen set van eigenschappen te achterhalen.

5. Tot slot

Aan dit project werkten mee: Eric van Balkum, Thomas Op de Coul, Kathrin Dentler, Nynke Kuipers, Ruben Schalk, Elena Slavco, Mirjam Verloop en Mari Wigham.

Heb je na het lezen van dit verslag nog vragen of ideeën, dan kun je het beste mailen met mirjam@podiumkunst.net.


podium
kunst
.net