OCLC | PICA

DAI werkprocessen

Onderzoeksgegevens in de persoonsnamenthesaurus

Auteur:	Martin van Muyen
Project:	DAI
Datum:	27-06-2006
Document naam:	DAI werkprocessen
Document status:	Definitief
Versie:	2.0
Opmerkingen:	Tot versie 1.2 o.d.t.: DAI formataanpassingen

DOCUMENT GESCHIEDENIS

Datum	Versie	Auteur	Opmerkingen
28-09-2005	0.1	MvM	Creatie: **DAI formataanpassingen**
07-11-2005	0.2	MvM	Wijzigingen in Inleiding en hoofdstuk 4 in rood; nieuw toegevoegd: hoofdstuk 3, 5, 6 en 7
16-11-2005	0.3	MvM	Hoofdstuk 3 en 7 aangepast na review TvdV en RK
25-11-2005	1.0	MvM	Besproken en geaccepteerd in DAI project bijeenkomst 25-11-2005; ongewijzigd behalve enige typo's
21-12-2005	1.1	MvM	Wijziging in 3.2.1 en 6 (Ophalen DAI); in 4.3 is veld Code-aanstelling is vervallen; appendices toegevoegd
26-01-2006	1.2	MvM	Arceringen verwijderd; opmerking over naamsvariant-en toegevoegd in 3.2.1
01-02-2006	1.3	MvM	Nieuwe titel: **DAI werkprocessen**
01-02-2006	1.3	MvM	Toegevoegd: 3.2.2 Export van thesauruswijzigingen; tekstuele aanpassingen in rood.
12-02-2006	1.4	MvM	Template voorbeelden en URL structuur Metis gegevens toegevoegd
22-02-2006	1.5	MvM	XML schema toegevoegd in Appendix 2; format aanpassingen en Appendix 4 in rood
05-04-2006	1.6	MvM	Herziene versie met wijzigingen in rood; Appendix III ingevoegd met drie template voorbeelden.
16-06-2006	1.7	MvM	Wijzigingen in XML schema in appendix 2; URL ipv script voor DAI EXPORT button in 3.2.1
27-06-2006	2.0	MvM	Ongewijzigde eindversie

INHOUD

1 Referenties

- Hans Schoonbrood – Tabellen – kolommen Metis voor DAI
- DAI PID – P1.1
- Format voor Persoonsnamen Thesaurus: OCLC PICA: dn010-persoonsnamenthesaurus

2 Inleiding

DAI is de afkorting van Digital Author Identification. De term is afkomstig uit het DARE project Orion waarin een plan was uitgewerkt om aan onderzoekers die verbonden zijn aan Nederlandse onderzoeksinstellingen een uniek nummer toe te kennen, een 'digital author identification'. In de database van GGC/NCC (ook bekend als 'de Pica database) wordt sinds lang een thesaurus onderhouden van persoonsnamen en personen. In deze thesaurus wordt voor een persoon een record aangemaakt, waarin voorkeursvorm, volledige naam en voorkomende naamsvarianten worden opgenomen. Aan zo'n record wordt een uniek nummer toegekend dat ppn (pica productienummer) genoemd wordt. Bibliotheken voegen namen van auteurs aan de thesaurus toe, wanneer er werken in de collectie worden opgenomen. Het is dan ook aannemelijk dat onderzoekers die publicaties op hun naam hebben staan, al in de thesaurus zijn opgenomen en daardoor in feite al een DAI hebben.

Het DAI project heeft tot doel een pilot uit te voeren in het gebruik van de Pica persoonsnamenthesaurus als basis voor de toekenning van DAI's aan onderzoekers. De pilot wordt uitgevoerd met de Metis onderzoek-databank van de RU Groningen. De Pica persoonsnamenthesaurus wordt tot nu toe uitsluitend gebruikt in het kader van de bouw van bibliotheekcatalogi. Bijgevolg zijn de huidige werkprocessen in het gebruik van de thesaurus afgestemd op het werkproces van catalogiseren, het maken van titelbeschrijvingen en het bouwen van catalogi.

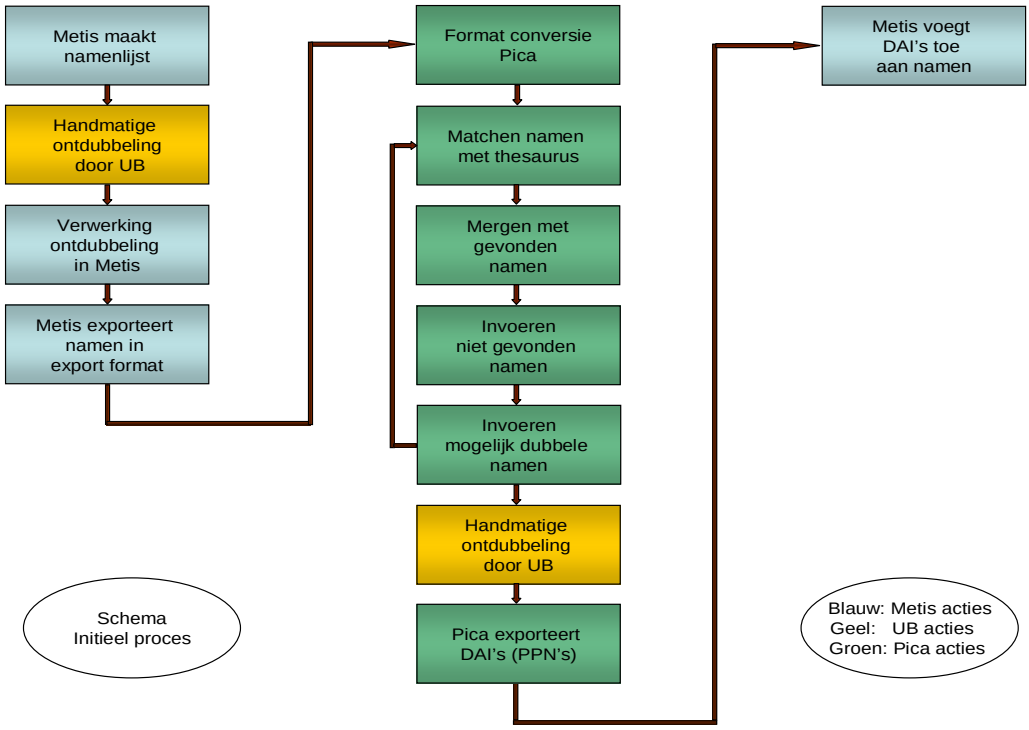
Het voorliggende document bevat een beschrijving van de werkprocessen bij het gebruik van de thesaurus ten behoeve van onderzoeksdatabanken en de specificatie van nieuwe velden en subvelden in het format van de persoonsnamenthesaurus die nodig zijn voor het vastleggen van onderzoekgegevens. De beschreven velden zijn afgestemd op de structuur en gebruik van de Metis onderzoeksdatabank van de Rijksuniversiteit Groningen. De gegevens worden vastgesteld in het kader van het DAI project en hebben tot doel normstellend te zijn voor alle onderzoeksdatabanken van de Nederlandse universiteiten.

3 Procesbeschrijving

In het pilot project zullen de persoonsgegevens afkomstig uit de Metis databank van de RU Groningen worden toegevoegd aan de persoonsnamenthesaurus. Na een initiële vulling dient een verwerkingsproces te worden ingericht dat wijzigingen die in Metis worden aangebracht automatisch in de thesaurus verwerkt. Tevens wordt een proces ingericht dat wijzigingen die in de thesaurus worden aangebracht kan exporteren naar een Metis bestand.

3.1 Initiële vulling

De initiële vulling van de persoonsnamenthesaurus met Metis gegevens zal als volgt plaats vinden:



Als eerste zal uit de Metis databank van de RUG een lijst worden samengesteld van namen die in de persoonsnamenthesaurus opgenomen zullen worden. Medewerkers van de UB Groningen zullen deze lijst controleren en aangeven welke namen dubbel zijn. Deze namen zullen in de Groningse Metis databank worden opgeschoond. Vervolgens worden de namen vanuit Metis geëxporteerd in het afgesproken export format (zie Appendix 1) en bij Pica op een FTP site geplaatst.

Bij Pica worden de gegevens verwerkt met de beschikbare batch invoer software. Allereerst worden de gegevens geconverteerd naar de overeengekomen datastructuur (Zie Nieuwe velden). Vervolgens worden overeenkomende namen in de thesaurus opgezocht. Daarbij zullen de matching criteria toegepast worden zoals beschreven in Matchen van persoonsnamen. Het matching proces leidt tot de volgende resultaten: er worden thesaurusrecords gevonden waarbij het duidelijk is dat het om dezelfde persoon gaat, er worden geen thesaurusrecords gevonden en er worden thesaurusrecords gevonden waarbij het niet duidelijk is dat het om dezelfde persoon gaat. Als deze laatste categorie vrij groot is, zal worden onderzocht of de matching criteria aangescherpt kunnen worden zodat deze namen nogmaals met de batch software verwerkt kunnen worden.

Uiteindelijk zal er een hoeveelheid namen overblijven die als 'mogelijk dubbel' gekenmerkt zal worden en met dit kenmerk zal worden ingevoerd. Deze categorie zal door deskundigen van de UB Groningen online bekeken worden en manueel worden ontdubbeld met de daarvoor beschikbare functies in de catalogiseer-client WInIBW.

Bij de gevonden thesaurus records zullen de Metis gegevens worden toegevoegd aan de reeds aanwezige gegevens. Als er geen thesaurus records worden gevonden, wordt een nieuw record aangemaakt op basis van de door de Metis databank aangeboden gegevens. (Zie ook Relatie tussen bestaande en nieuwe velden.)

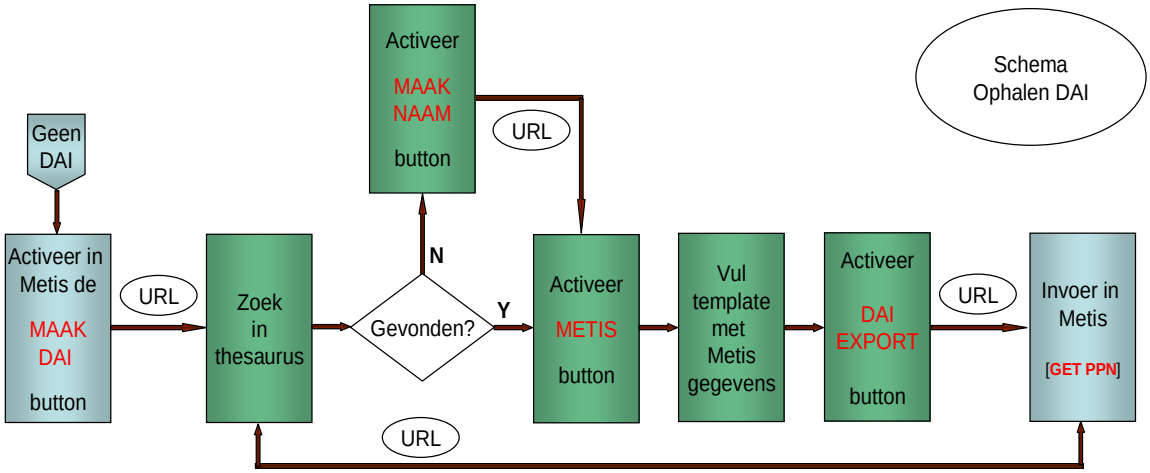
Nadat de Metis gegevens in de thesaurus zijn verwerkt, wordt door Pica een lijst geproduceerd van alle namen waaraan onderzoeksgegevens zijn toegevoegd. De lijst zal voor elke naam alle naams- en onderzoeksgegevens bevatten in een XML structuur. Op basis van deze lijst kunnen de PPN's (Pica productienummers) als DAI's aan de gegevens in de Metis databank worden toegevoegd. Desgewenst kan deze lijst twee keer worden gemaakt: de eerste keer voor alle namen die gevonden of ingevoerd zijn, de tweede keer van de namen die door deskundigen van de UB zijn ontdubbeld. Als een Metis databank de aldaar opgeslagen naamsgegevens wil verrijken met de naamsgegevens uit de thesaurus kan hiervoor dezelfde lijst gebruikt worden. Zie ook Export van thesauruswijzigingen en Appendix 2.

3.2 Dagelijks gebruik

Na de initiële vulling zullen de Metis gegevens primair in de Metis databank worden onderhouden: als een onderzoeker al een DAI heeft, zullen wijzigingen in de Metis databank worden aangebracht en zal via een periodieke batch upload de relevante gegevens aan de betreffende thesaurusrecords worden toegevoegd. Alleen voor nieuwe onderzoekers zal eerst in de thesaurus gezocht worden, om een DAI op te halen.

3.2.1 Ophalen DAI

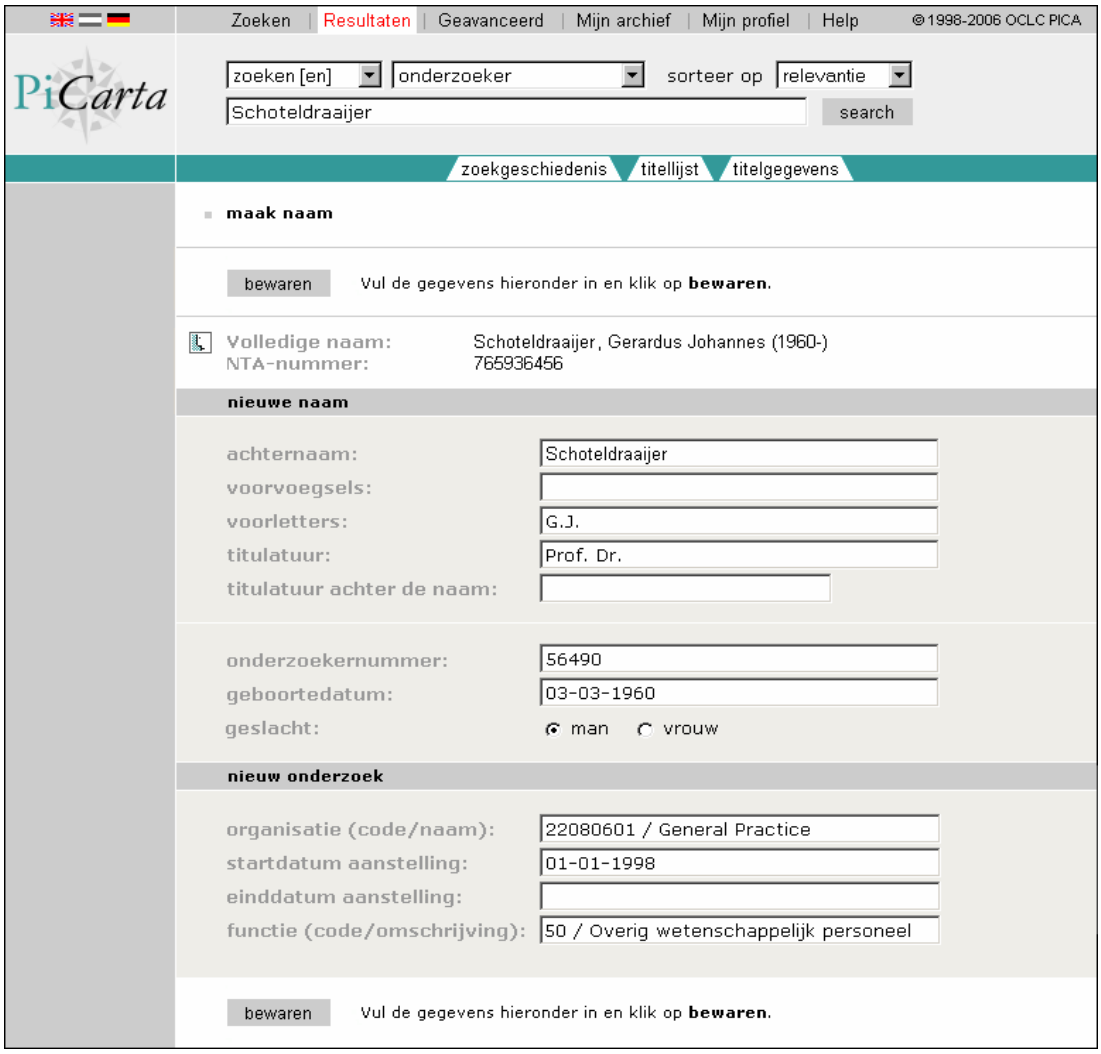
Voor onderzoekers zonder DAI geldt het onderstaande procesverloop:



Wanneer in de Metis databank een naam niet aanwezig is, dan wel gevonden wordt zonder een DAI (naam afkomstig uit universitaire administratie), moet een DAI worden 'aangemaakt'. Daartoe wordt in de Metis interface een 'MAAK DAI' button aangeboden. Wanneer de button wordt geactiveerd, wordt een URL samengesteld die er voor zorgt dat de gebruiker automatisch inlogt in het GGC en automatisch een zoekactie uitvoert naar in de persoonsnamenthesaurus aanwezige onderzoekers (zie ook Benodigde URL's). Indien er geen onderzoeker wordt gevonden, dient de gebruiker de zoekactie te herhalen voor alle in de thesaurus aanwezige namen door het sleuteltype van de zoekvraag te wijzigen.

Als na een tweede zoekactie nog geen naam is gevonden, dient de gebruiker de 'MAAK NAAM' button te klikken. Deze button zal aan daartoe bevoegde gebruikers worden aangeboden op de zoek template van het Pica systeem. Na activering van deze button wordt een URL samengesteld met als inhoud de naam waarmee gezocht is, uitgebreid met een aantal standaard gegevens die benodigd zijn om automatisch een nieuw thesaurusrecord met onderzoeksgegevens aan te maken (zie ook Relatie tussen bestaande en nieuwe velden).

Na een succesvolle zoekactie of na het aanmaken van een nieuw thesaurusrecord wordt de naam getoond in een gelabelde presentatie. Op dat template zal aan daartoe bevoegde gebruikers de METIS button worden getoond voor het automatisch laten toevoegen van de Metis gegevens, beschreven in Nieuwe velden. Na activeren van de METIS button verschijnt het onderstaande scherm:



Invul template

Nadat de door het systeem ingevulde velden door de gebruiker zijn gecontroleerd en zonodig verbeterd en door het systeem zijn geaccepteerd[1], worden alle gegevens nogmaals getoond met een 'DAI EXPORT" button die een URL activeert om het ppn naar Metis te transporteren. Metis kan het DAI-nummer toevoegen

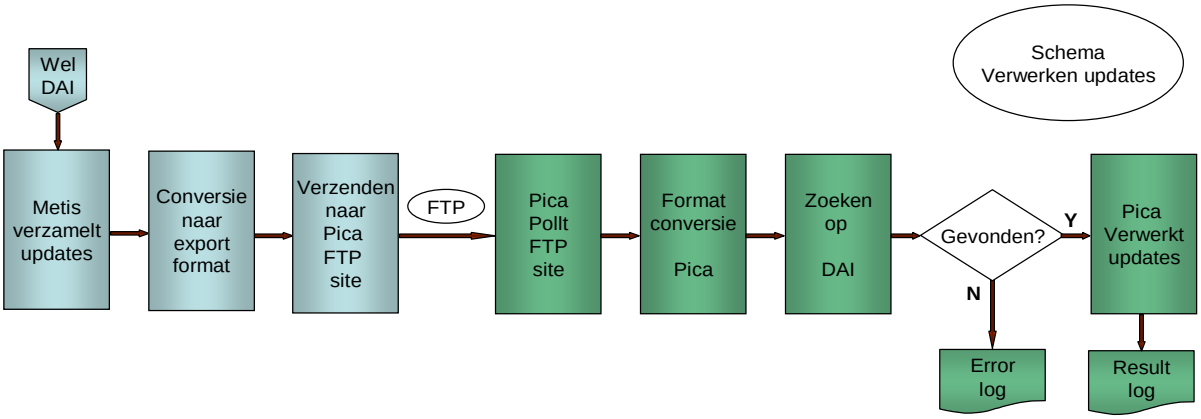
[1] Merk op dat het invul-template geen velden bevat voor mogelijke naamsvarianten; er wordt vanuit gegaan dat deze bij (nieuwe) onderzoekers die nog geen DAI hebben, nog niet voorkomen. Mogelijke naamsvarianten kunnen in later stadium met behulp van het proces Ophalen van Metis wijzigingen worden toegevoegd. Op de templates is de term DAI vervangen door NTA (Nederlandse Thesaurus voor Auterursnamen).

aan de reeds aanwezige naamsgegevens of, indien nodig de zojuist ingevoerde gegevens uit de thesaurus opgehalen door een 'GET PPN' URL te sturen. Afhankelijk van de URL-definitie worden de gegevens in een gelabelde of in een XML presentatie getoond. Onderstaande voorbeeld toont een gelabelde presentatie.

Gelabelde presentatie met onderzoeksgegevens

3.2.2 Ophalen van Metis wijzigingen

Voor wijzigingen die in het Metis bestand worden aangebracht aan de gegevens van onderzoekers die al een DAI hebben geldt het volgende procesverloop:



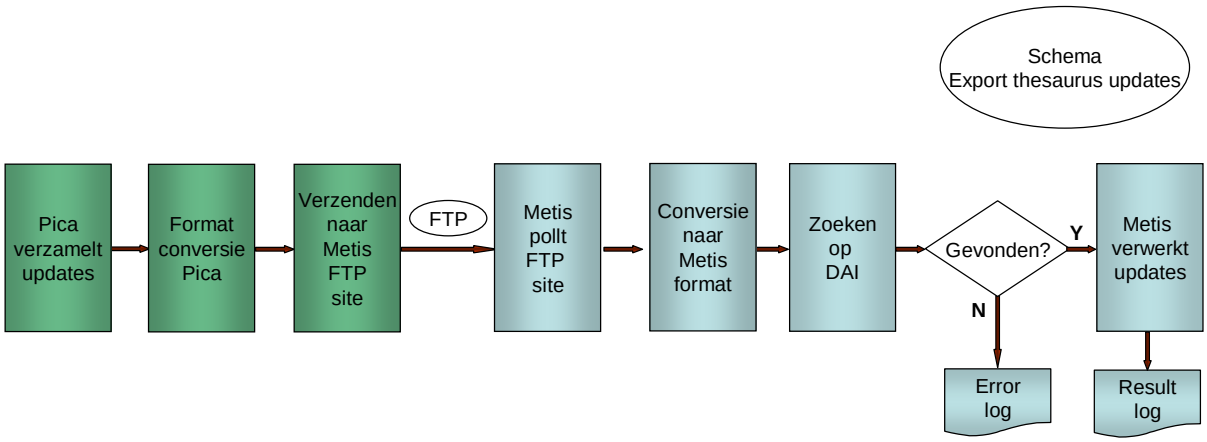
In Metis worden periodiek, bijvoorbeeld één keer per dag de gewijzigde gegevens van onderzoekers met een DAI verzameld en geconverteerd naar het overeengekomen export format. De geconverteerde gegevens worden op een FTP site van Pica geplaatst. Bij Pica is een polling mechanisme actief dat alle geadministreerde FTP sites controleert op aanwezigheid van nieuwe FTP files. Gevonden gegevens afkomstig uit

Metis worden geconverteerd naar het Pica+ format en opgezocht in de thesaurus. Voor dit 'matchen' wordt het PPN / DAI gebruikt. Theoretisch kan het voorkomen dat het gezochte thesaurusrecord niet meer aanwezig is; in dat geval wordt een foutmelding geschreven in een logfile.

Bij het verwerken van de updates worden bij elk gevonden thesaurusrecords per universiteit alle aanwezige Metisnamen vervangen door de nieuwe namen en worden alle aanwezige onderzoeksblokken vervangen door de nieuwe. Daarbij blijven de reeds toegekende onderzoeksblok productienummers (epn's) zoveel mogelijk gehandhaafd. [Bijv.: aan een thesaurusrecord waren drie onderzoeksblokken toegevoegd; er wordt nu een update aangeboden met twee onderzoeksblokken: de eerste twee blokken worden gemuteerd, het derde wordt verwijderd.]
Overwogen kan worden om, bij voorkeur in tweede instantie, na een succesvol verlopen pilot, het FTP proces te vervangen door een web service die het mogelijk maakt om een mutatie in Metis direct naar de thesaurus te sturen.

3.2.3 Export van thesauruswijzigingen

Voor Metis bestanden die gebruik willen maken van de naamsgegevens uit de thesaurus wordt een export proces ingericht dat wijzigingen in thesaurusnamen automatisch distribueert naar de betreffende Metis bestanden. Dit proces is gelijk aan het Ophalen van Metis wijzigingen, maar met tegenovergestelde actoren:



Batch jobs bij Pica die gedefinieerd en geactiveerd worden met de Export module van het CBS systeem, controleren dagelijks in de CBS log file of er updates zijn voor persoonsnamen waaraan Metis gegevens gekoppeld zijn. Dit proces selecteert niet alleen wijzigingen die door de eigen instelling zijn ingevoerd, maar ook wijzigingen die door andere instellingen, zoals de Koninklijke Bibliotheek, zijn aangebracht aan namen waaraan Metis gegevens van de eigen instelling zijn gekoppeld. Als default worden alleen de namen geselecteerd waaraan Metis gegevens van de eigen instelling gekoppeld zijn. Na de pilot dient dus voor elke onderzoeksinstelling die de naamsgegevens van de thesaurus wenst vast te leggen in het eigen Metis bestand een batch job gedefinieerd te worden.

De geselecteerde namen worden geconverteerd naar het XML format dat is gespecificeerd in Appendix 2. Alle velden van het thesaurus record en alle aangehechte onderzoeksblokken van de betreffende onderzoeksinstelling worden uitgevoerd. De geselecteerde records worden door Pica verzonden naar de FTP site behorend bij het betreffende Metis bestand, dat door Metis gecontroleerd dient te worden op de aanwezigheid van nieuwe FTP files. Vervolgens converteert Metis de gegevens naar de eigen datastructuur, zoekt de corresponderende records op in het Metis bestand met behulp van het DAI van het geëxporteerde record en verwerkt de wijzigingen in het Metis bestand. Vooralsnog wordt ervan uitgegaan dat niet-gevonden namen in een foutenlijst worden bijgehouden voor manuele controle.

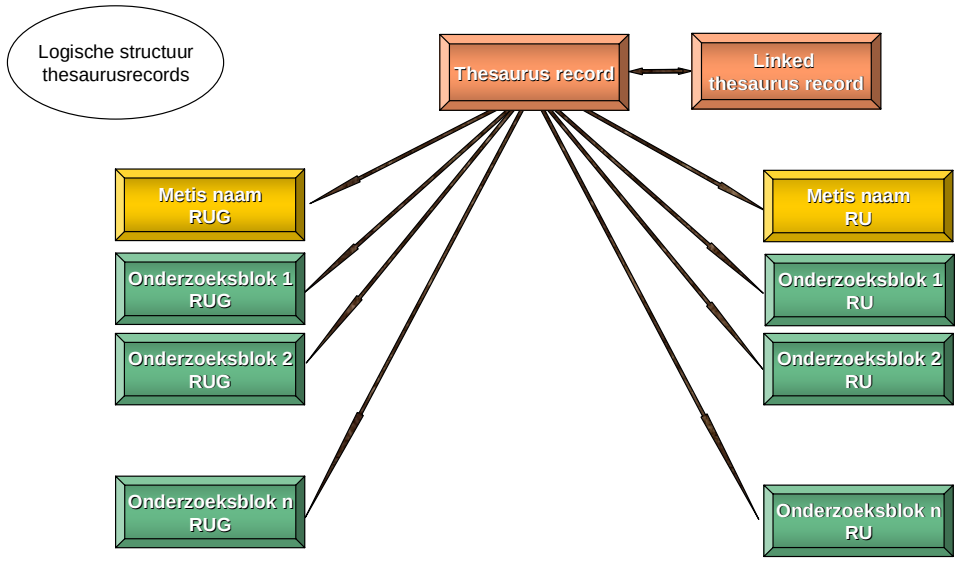
Gebruik van de CBS log file voor de distributie van wijzigingen heeft als voordeel dat voor het exporteren van namen na de initiële vulling en voor de export van latere wijzigingen hetzelfde proces gebruikt kan worden; ook de wijzigingen aan thesaurusrecords die het gevolg zijn van de initiële vulling worden immers gelogd.

Een speciale vorm van wijzigingen hebben te maken met opschoonacties die het gevolg kunnen zijn van de initiële vulling maar die ook daarna kunnen voorkomen, wanneer ontdekt wordt dat twee thesaurusrecords moeten worden samengevoegd of wanneer één thesaurusrecord moet worden gesplitst. Dit kan leiden tot het overhevelen van de aangehechte onderzoeksblokken naar een ander thesaurusrecord. In die gevallen krijgt een onderzoeker dus een nieuwe DAI. Samenvoegen en splitsen van naamsrecords is een actie die alleen door bibliothecarissen mag worden uitgevoerd met de catalogiseer-client WinIBW. Voor consistentie van de Metis bestanden is het noodzakelijk dat als een onderzoeksblok wordt overgeheveld van het ene naamsrecord naar het andere en een onderzoeker daardoor een nieuwe DAI krijgt, het oude DAI in het record van de nieuwe DAI wordt opgenomen. Het betreffende veld is beschreven in hoofdstuk 4.2.
Appendix 4 geeft een procedurebeschrijving voor het splitsen en samenvoegen van namen in de persoons-namenthesaurus. De daarin genoemde WinIBW scripts zullen ter beschikking worden gesteld door Pica.

4 Nieuwe velden

Omdat onderzoekers verbonden kunnen zijn aan meerdere vakgroepen, faculteiten en universiteiten wordt bij het vaststellen van de structuur gebruik gemaakt van een technische mogelijkheid die tot nu toe uitsluitend voor metadata wordt gebruikt: de mogelijkheid om gegevens op 'exemplaarniveau' vast te leggen. Bij metadata, zoals de titelgegevens van boeken, wordt het exemplaarniveau gebruikt om een signatuur en uitleencode of, bij tijdschriften, de aanwezige jaargangen vast te leggen; bij persoonsnamen kan het worden gebruikt als een 'onderzoeksblok' dat herhaald wordt voor elk werkverband van een onderzoeker. Strekken de werkverbanden zich uit over meerdere universiteiten dan worden de werkverbanden per universiteit vastgelegd, waarbij altijd een koppeling wordt gemaakt naar dezelfde persoonsnaam.

Omdat de Metis gegevens primair in de Metis databanken zullen worden onderhouden, worden de uit Metis afkomstige naamsgegevens apart gehouden van de huidige velden van de persoonsnamenthesaurus. Ook hiervoor wordt gebruikt gemaakt van een aan titelbeschrijvingen ontleende veldstructuur: de Metis naamgegevens zullen worden vastgelegd in 'lokale' velden. Wanneer een onderzoeker aan meerdere universiteiten verbonden is en dus in meerdere Metis bestanden voorkomt en dus in meerdere Metis bestanden gewijzigd kan worden, zal voor elke universiteit een apart lokaal blok beschikbaar zijn. Eén en ander leidt tot de onderstaande recordstructuur:



4.1 Velden op gemeenschappelijk niveau

De onderzoeksgegevens hebben slechts in beperkte mate invloed op de gemeenschappelijke gegevens in de persoonsnamenthesaurus: er wordt een veld door het systeem gegenereerd als aan een thesaurusrecord onderzoeksgegevens worden toegevoegd en er wordt een veld toegevoegd dat wordt gebruikt wanneer bij de initiële vulling vanuit onderzoeksbestanden mogelijk dubbele namen worden ingevoerd.

Pica+	Thesaurus	Metis	Omschrijving
001@	-	-	ILN bit string
Toelichting: 1. Wordt per universiteit geactiveerd als een eerste onderzoeksblok wordt toegevoegd en wordt gede-			

activeerd als het laatste blok wordt verwijderd.

2. Wordt in het thesaurus format niet getoond
3. Kan in overige presentaties worden omgezet in de naam van de instelling
4. Activering en de-activering worden in het GGC gelogd als update
5. Speelt bij thesaurus records (vooralsnog) geen rol bij zoeken en filteren

Pica+	Thesaurus	Metis	Omschrijving
038L	796	-	DAI – PPN van voorkeursnaam
$a			Bron van nieuwe naam
$b			Status van nieuwe naam
$x			Gelijkheidswaarde van beide namen
$9	! … !	-	PPN van voorkeursnaam

Toelichting:

Dit veld wordt aan en nieuwe naam toegevoegd bij batch invoer, wanneer bij het matchen een naam gevonden wordt met een gelijkheidswaarde die ligt tussen de waarden die gedefinieerd zijn voor samenvoegen en voor nieuwe invoer; de ingevoerde naam is dus mogelijk dubbel en veld 038L bevat naast enige administratieve informatie het PPN van de gevonden naam. Zie ook Matchen van persoonsnamen.

4.2 Velden op lokaal niveau

Op het nieuw te definiëren lokale niveau worden alle Metis naamsgegevens en het lokale Metis onderzoekernummer vastgelegd. In beginsel zullen deze velden zichtbaar zijn voor de bibliotheekmedewerkers van de eigen universiteit, die de thesaurus gebruiken voor de catalogusbouw; indien echter nodig kunnen de lokale velden met behulp van een filtertabel verborgen worden. Gebruikers van de persoonsnamen thesaurus van andere universiteiten kunnen de onderzoeksgegevens alleen zien als men bevoegd is lokale en exemplaargegevens van andere bibliotheken te bekijken. Invoer en mutatie zijn uitsluitend voorbehouden aan medewerkers van de eigen universiteit.

In de lokale gegevens wordt tevens het veld ‘Oud PPN – Oud DAI’ opgenomen dat gebruikt wordt als thesaurusrecords worden samengevoegd of gesplitst. Strikt genomen hoort dit veld niet op lokaal niveau, maar in het huidige gebruik van de persoonsnamenthesaurus is aan vastlegging van het oude PPN bij samenvoeging en splitsing (tot nu toe) geen behoefte. Het veld is echter voor de DAI administratie in Metis bestanden zo belangrijk dat het voor onderzoekers wordt vastgelegd op lokaal niveau.

Toevoegen van lokale velden, dus van Metis naamsgegevens, is alleen mogelijk als tevens een onderzoeksblok wordt toegevoegd.

Pica+	Thesaurus	Metis	Omschrijving
103M	A01	ONDERZOEKERNUMMER	Lokaal onderzoekernummer
$0			Nummer van maximaal 10 posities

Toelichting:

1. Dit veld is benodigd om DAI’s (ppn’s) te kunnen opnemen in de Metis databank; na afloop van het project kan bekeken worden of dit nummer opgeheven kan worden.

2. Het veld zal door het Pica systeem niet inhoudelijk worden gecontroleerd.
3. Het veld zal worden geïndexeerd.

Pica+	Thesaurus	Metis	Omschrijving
103Z	A99	Tbd	Oud PPN – oud DAI
$0			Oud PPN

Toelichting:

Kenmerk 103Z wordt met behulp van WinIBW scripts toegevoegd indien onderzoeksgegevens worden overgeheveld van een verwijderd thesaurusrecord naar record dat door de bibliothecaris is aangemerkt als het voorkeursrecord. Deze actie kan alleen uitgevoerd worden door een bibliothecaris die is verbonden aan dezelfde onderzoeksinstelling.

Pica+	Thesaurus	Metis	Omschrijving
128A	A10	-	Metis naamsvorm met initialen
$e	#...#	TT	Titulatuur
$d		VLT	Voorletters
$c	/…	VV	Voorvoegsel
$a	@...	NAAM	Eigennaam, achternaam
$f	(…)	-	Titulatuur achter de naam
		VOORKEUR	Zie toelichting punt 1

Toelichting:

1. Het Metis veld VOORKEUR wordt gebruikt in de conversie naar Pica+:
   - indien dit element aanwezig is, wordt kenmerk 128A aangemaakt;
   - indien dit veld niet aanwezig is, worden de naamsgegevens vertaald naar 128@ (zie volgende tabel)
2. De inhoud van kenmerk 128A wordt drie keer geïndexeerd met de voor persoonsnamen gebruikelijke indexroutine:
   - Eén keer als een lokale index, alleen toegankelijk voor de eigen bibliotheek / universiteit
   - Twee keer als een algemeen doorzoekbare index (ILN=0). (Deze index zal na de implementatie van CBS4 v3.2 beschikbaar komen.)
0. Er zullen twee algemeen doorzoekbare indexen gemaakt: één met het voor persoonsnamen gebruikelijke indextype en één met een nieuw indextype. Wanneer bij zoeken van dit nieuwe indextype gebruik wordt gemaakt, zoekt men impliciet in de sub-thesaurus van onderzoekers; in dummy commandotaal:
   - <zoek onderzoeker> zoekt binnen de thesaurus records met een 128A
   - <zoek persoon> zoekt binnen de gehele thesaurus.
1. Subveld $f is gedefinieerd voor toekomstig gebruik; het zal niet worden gevuld vanuit de initiële Metis invoer.

Pica+	Thesaurus	Metis	Omschrijving
128B	A11	-	Naam met uitgeschreven voornaam/namen
$e	#...#	TT	Titulatuur
$d		VLT	Voornaam/namen

$c	/…	VV	Voorvoegsel
$a	@...	NAAM	Eigennaam, achternaam
$f	(…)	-	Titulatuur achter de naam

Toelichting:

0. In Metis wordt het veld voornaam/voornamen nog niet ondersteund; daarom wordt het in de pilot direct aangeleverd uit de personeelsadministratie van de RUG in een aparte tabelbestaande uit onderzoeker-nummers en voornamen. Op basis van de onderzoekernummers wordt bij de conversie een volledige 128B samengesteld met voornaam/namen, achternaam, voorvoegsel en titel.
2. Zie verder de toelichting bij 128A.

Pica+	Thesaurus	Metis	Omschrijving
128@	A20	-	Metis naamsvorm – niet voorkeursnaam
$d		VLT	Voorletters, roepnaam
$c	/…	VV	Voorvoegsel
$a	@...	NAAM	Eigennaam, achternaam

Toelichting:

1. De Metis naamsgegevens worden geconverteerd naar 128@ indien het element VORRKEUR ontbreekt.
2. Er is afgesproken dat in niet-voorkeursnamen geen titulatuur wordt opgenomen.
3. De inhoud van kenmerk 128@ wordt op dezelfde wijze geïndexeerd als kenmerk 128A.

Pica+	Thesaurus	Metis	Omschrijving
132A	A30		Geboortedatum / leefjaren
$a		GEB_DATUM	Leefjaren
$b	=	GESL	Geslacht: 1 positie: M of V

Toelichting:

1. Dit veld is optioneel; indien aanwezig is minimaal één subveld verplicht; in het thesaurus format begint dit veld dus met een '=' als alleen de geslachtsaanduiding aanwezig is.
2. Het geboortejaar wordt bij offline matchen gebruikt om te bepalen of dezelfde naamsvormen verschillend zijn.

4.3 Velden van het onderzoeksblok

De velden die tot het onderzoeksblok gerekend worden, zijn gedefinieerd op niveau 2 van de Pica+ data-structuur. Voor onderzoekers met meerdere werkverbanden, wordt per werkverband een onderzoeksblok gedefinieerd. In beginsel zullen deze velden zichtbaar zijn voor de bibliotheekmedewerkers van de eigen universiteit, die de thesaurus gebruiken voor de catalogusbouw; indien echter nodig kunnen ook de exemplaargebonden velden met behulp van een filtertabel verborgen worden. Gebruikers van de persoons-namenthesaurus van andere universiteiten kunnen de onderzoeksgegevens alleen zien als men bevoegd is lokale en exemplaargegevens van andere bibliotheken te bekijken. Invoer en mutatie zijn uitsluitend voorbehouden aan medewerkers van de eigen universiteit.

Pica+	Thesaurus	Metis	Omschrijving
208@	E01		Selectiesleutel
$a			Datum, wordt door systeem gegenereerd
$b			Selectiesleutel; zie toelichting

Toelichting:

1. In de structuur van onderzoeksblokken (exemplaarblokken) is veld 208@ noodzakelijk. Het veld wordt in de DAI web interface niet getoond (zie Ophalen DAI). Indien niet aanwezig wordt het door het systeem automatisch toegevoegd onder de volgende condities:
   - de server behorend bij de web interface genereert veld 208@ subveld $b met als inhoud een kleine letter 'p'.
   - tevens wordt door de catalogiseer server subveld $a toegevoegd met als inhoud de systeemdatum.
   - indien veld 208@ reeds aanwezig is in en onderzoeksblok kan dit veld via de DAI web interface niet gewijzigd worden; wijzigingen zijn dan alleen mogelijk met WinIBW.
2. Er wordt een gecombineerde index op de inhoud van beide subvelden: datum + code. Deze ingang is per bibliotheek (ILN) afzoekbaar.

Pica+	Thesaurus	Metis	Omschrijving
229A	B10		Organisatie behorend bij het dienstverband
$B		CODE_ORGANISATIE	Code van de organisatie waarbij de onderzoeker een aanstelling heeft
$a			Bij code behorende naam

Toelichting:

1. Dit veld is herhaalbaar; thans komt het in Metis per dienstverband (onderzoeksblok) maximaal drie keer voor; in de toekomst zal deze beperking worden opgeheven.
2. De bij een code behorende naam zal worden meegeleverd. Codes en namen zijn per Metis database verschillend; daarom worden codes in presentaties niet met behulp van tabellen geëxpandeerd.
3. In eindgebruikerpresentaties wordt alleen de inhoud van $a getoond; in het catalogiseerformat wordt de inhoud van $B achter de inhoud van $a getoond. .
4. Beide subvelden worden geïndexeerd:
   - de code in $B wordt geïndexeerd per bibliotheek (ILN)
   - de naam in $a wordt zowel per bibliotheek als algemeen doorzoekbaar (ILN=0) geïndexeerd.

   In beide gevallen worden er twee indexen gemaakt: een index op woordbasis en een index op de gehele inhoud van $a. (De index met ILN=0 zal gemaakt worden na implementatie van CBS4 v3.2)

Pica+	Thesaurus	Metis	Omschrijving
232A	B20		Start- en einddatum aanstelling
$a		BEGIN_PERIODE	Startdatum aanstelling
$b		EIND_PERIODE	Einddatum aanstelling

Toelichting:

1. Dit veld wordt niet geïndexeerd.
2. De structuur van begin- en einddatum wordt niet gecontroleerd.

Pica+	Thesaurus	Metis	Omschrijving
232B	B21		Codering en omschrijving functie
$a		CODE_FUNCTIE	Code van de functie
$b			Bij code behorende omschrijving
Toelichting: 1. Dit veld wordt niet geïndexeerd.			
Pica+	**Thesaurus**	**Metis**	**Omschrijving**
237A	B50	-	Veld voor opmerkingen / toelichtingen
$a		-	Opmerking / toelichting
Toelichting: 1. Dit veld is herhaalbaar 2. Dit veld wordt niet geïndexeerd. 3. Het veld wordt niet aangeleverd bij de initiële invoer vanuit Metis.			
Pica+	**Thesaurus**	**Metis**	**Omschrijving**
203@	B90		Onderzoeksblok productienummer (epn)
$0			Nummer: 9-10 posities, incl. check digit
Toelichting: 1. Dit veld wordt geïndexeerd.			
Pica+	**Thesaurus**	**Metis**	**Omschrijving**
201B	B95		Datum en tijdstip laatste mutatie van onderzoeksblok
$0			Datum: dd-mm-jj
$t			Tijdstip: uu-mm-ss-mmm
Toelichting: 1. Dit veld wordt niet geïndexeerd			

5 Relatie tussen bestaande en nieuwe velden

Zoals beschreven in hoofdstuk 4 worden alle uit Metis afkomstige velden vastgelegd op lokaal en exemplaar niveau. Daarmee wordt bereikt dat de twee toepassingsgebieden van de Pica persoonsnamenthesaurus elkaar niet of nauwelijks kunnen verstoren, hoezeer ook verschillend in gebruik. Er is echter één punt waarop gebruik kan interfereren: wanneer vanuit Metis gegevens moeten worden toegevoegd aan een naam die nog niet in de thesaurus aanwezig is. In dat geval dient een thesaurusrecord te worden aangemaakt ten einde de gewenste lokale en exemplaargegevens te kunnen toevoegen. Deze situatie kan zowel bij initiële vulling als bij het dagelijks gebruik optreden.

Initiële vulling:

Als bij de initiële vulling geen naam wordt gevonden in de Persoonsnamenthesaurus wordt een nieuw thesaurusrecord aangemaakt op basis van de Metis naamsgegevens:

- 128A [kenmerk A10] wordt gekopieerd naar 028A [kenmerk 100], met dezelfde subvelden.
- 128B [kenmerk A11] wordt gekopieerd naar 028B [kenmerk 110], met dezelfde subvelden.
- 132A [kenmerk A30] wordt gekopieerd naar 032A [kenmerk 300], met hetzelfde subveld.

Tevens wordt 002@ [kenmerk 005] toegevoegd met als inhoud Tpx.
Vervolgens worden aan dit nieuwe thesaurusrecord de lokale en exemplaargegevens toegevoegd.

Dagelijks gebruik:

Bij het online ophalen van DAI's kan het voorkomen dat een naam nog niet in de persoonsnamenthesaurus aanwezig is. De gebruiker kan dan op een MAAK NAAM button klikken waarna het Pica systeem ervoor zorgt dat op basis van de zoekstring een naam wordt gemaakt bestaande uit de volgende Pica+ velden:

- 002@ [005] Tpx
- 028A [100] met als inhoud de uit de zoekstring gekopieerde naam. Daarbij wordt tekst voor een komma vertaald naar het subveld voor de achternaam ($a) en tekst na de komma naar het subveld voor de voornaam ($d). Eventuele voorvoegsels worden als deel van de voornaam behandeld. Vooralsnog wordt ervan uitgegaan dat er één zoekstring wordt gebruikt; er zal bij het aanmaken van een thesaurusrecord dus geen rekening gehouden worden met namen met initialen en namen met uitgeschreven voornaam/ namen.

Nadat op deze wijze een naam aan de thesaurus is toegevoegd, kan de gebruiker doorgaan met het online verwerken van de Metis gegevens.

6 Benodigde URL's

In het dagelijkse gebruik van de persoonsnamenthesaurus in relatie met Metis zijn vier URL's en een java script voorzien om de procesgang te vereenvoudigen:

1. MAAK DAI

De mogelijkheid om deze URL te activeren wordt geboden in de Metis interface indien een gebruiker die toegang heeft tot de persoonsnamenthesaurus een naam zonder DAI geselecteerd of aangemaakt heeft. Met behulp van deze URL wordt een gebruiker in het GGC ingelogd en wordt een zoekactie gedaan naar de aanwezigheid van een onderzoeker. Daartoe bevat deze URL de volgende gegevens:

- Usernummer / wachtwoord dat gebruikt moet worden voor het inloggen. Het usernummer is gekoppeld aan de UB Groningen en krijgt de bevoegdheden die nodig zijn om te kunnen zoeken in de thesaurus en om nieuwe persoonsnamenrecords te kunnen maken en om lokale en exemplaargegevens te kunnen toevoegen (in het verborgen Pica+ format).
- Inperking van de zoekactie tot het Record type Thesaurus
- Gebruik van de standaard zoeksleutel <onderzoeker>
  Het door Metis te ontwikkelen script dat deze basisgegevens ophaalt, dient achter het sleuteltype de naamsgegevens te plakken in de structuur <achternaam komma voornaam>.
- Naams- en onderzoeksvelden, conform de definitie van hoofdstuk 4.2 en 4.3 in de URL-structuur beschreven in Appendix 1.

2. MAAK NAAM

De Maak NAAM URL wordt gebruikt om bevoegde gebruikers de mogelijkheid te bieden om na een zoekactie een naam in te voeren. De URL bevat de volgende elementen:

- Invoer commando
- De Pica+ velden 002@ $0Tpa; 028A $a<naam>$d<voornaam>. De inhoud van <naam> en <voornaam> worden uit de zoekstring gekopieerd.

3. METIS

De METIS URL wordt gebruikt om bevoegde gebruikers de mogelijkheid te bieden lokale en exemplaargegevens aan een gevonden naam toe te voegen. De bijbehorende button wordt alleen getoond in volledige presentatie van persoonsnamen. Na activering wordt de template getoond waarop de Metis naamsgegevens en onderzoeksgegevens kunnen worden toegevoegd en gemuteerd.

4. DAI EXPORT

De DAI EXPORT button wordt uitsluitend getoond op de template met de Metis gegevens. Na activering wordt een java script geactiveerd dat het PPN (DAI) op het klembord plaatst en het thesaurusvenster sluit. Op het Metis scherm kan de gebruiker vervolgens het PPN in het daartoe bestemde veld plakken.

5. GET PPN

De GET PPN button kan in Metis omgeving gebruikt worden om het complete thesaurusrecord met toegevoegde lokale en exemplaargebonden gegevens op te halen uit de thesaurus zodat deze in het Metis bestand verwerkt kan worden als een nieuwe invoer of als een update, al naar gelang de lokale situatie.

Desgewenst kunnen bovenstaande URL's als Open-URL's of als 'free-format' URL's worden geadministreerd en onderhouden in de Admin module bij Pica. In het kader van het pilot project zal Pica de benodigde URL's vastleggen en indien nodig aanpassen. Het java script achter de DAI export button wordt direct aan de betreffende button gekoppeld en is dus niet via een Admin module te wijzigen.

7 Matchen van persoonsnamen

Omdat alle uit Metis afkomstige gegevens worden geconverteerd naar lokale en exemplaargebonden velden in het Pica format, zal een initiële vulling altijd per universiteit plaats vinden, ook als in de toekomst meerdere universiteiten tegelijk hun Metis gegevens met de persoonsnamenthesaurus willen koppelen. Bij een initiële vulling worden de uit Metis geëxporteerde naamsgegevens gebruikt om overeenkomende namen in de persoonsnamenthesaurus op te zoeken. In de thesaurus zijn persoonsnamen als volgt geïndexeerd:

- Uitsluitend kleine letters
- UTF8 tekenset, met indexering van diakritische tekens:
  - Indexering van Franse accenten
  - Indexering van umlaut – geen vertaling naar grondletter + e
  - Indexering van Nederlandse IJ als i+j
- De structuur van de index is: achternaam, komma, voornaam/namen
- Voorvoegsels worden in de indexen achter de laatste voornaam opgenomen (als 'laatste voornaam').

Bij een initiële vulling worden de Metis gegevens eerst geconverteerd naar Pica kenmerken en worden daarna uit deze velden zoekingangen gegenereerd volgens dezelfde conventies als bij de indexen (kleine letters, UTF8, dezelfde behandeling van diakrieten etc.). Vervolgens wordt getracht de bij de Metis namen behorende thesaurusrecords te vinden. Dit zal in een aantal losse stappen gedaan worden, waarbij na elke stap de nog overblijvende namen geanalyseerd zullen worden zodat de volgende stap een zo groot mogelijk rendement zal opleveren.

Bij elke stap wordt automatische aanwezigheidscontrole toegepast. Bij deze controle worden de gelijkheidswaarden berekend tussen de in te voeren namen en de gevonden namen en worden onder- en bovengrenzen vastgesteld:

- namen die onder de ondergrens blijven (bijvoorbeeld 0.10) worden als nieuwe namen ingevoerd
- bij namen die boven de bovengrens komen (bijvoorbeeld 0.75) worden de onderzoeksgegevens aan de gevonden naam toegevoegd

Namen met een gelijkheidswaarde tussen de onder- en bovengrens zullen in eerste instantie apart gehouden voor nadere analyse om in een volgende stap de aanwezigheidscontrole te verfijnen.

Vanaf de eerste stap zal het analyseren van de leefjaren een belangrijke rol spelen. Dit veld wordt in de thesaurus niet geïndexeerd en speelt dus pas een rol als er thesaurusrecords worden gevonden. Daarbij doen zich de volgende situaties voor:

- geen leefjaren in beide records, d.w.z. in het Metis record en het eerst gevonden thesaurus record
- gelijke leefjaren in beide records
- verschillende leefjaren in beide records
- leefjaren in één van beide records.

Deze situaties kunnen tot uitdrukking gebracht worden in het berekenen van de gelijkheidswaarden. Indien nodig kunnen voor volgende stappen deze berekeningen gewijzigd worden; ook kunnen speciale indexen gebouwd worden om de trefkans te vergroten, bijvoorbeeld door bij voornamen alleen de eerste letter te gebruiken om beter met initialen te kunnen matchen.

Een eerste run kan bijvoorbeeld beperkt worden tot Metis namen die slechts één treffer opleveren in de thesaurus. Indien de gevonden thesaurusrecords dezelfde leefjaren hebben als de Metis records, worden de Metis velden als lokale en exemplaargebonden velden toegevoegd. Alle overige namen worden in deze eerste run nog niet verwerkt. Na evaluatie van de overgebleven namen kan een tweede run gedraaid worden waarbij, bijvoorbeeld alle namen die geen treffer hebben opgeleverd met behulp van speciale indexen nogmaals worden opgezocht. In een derde run zouden niet-gevonden namen en namen met een gelijk-

heidswaarde kleiner dan (bijvoorbeeld) 0.20 in de thesaurus kunnen worden opgenomen. Een volgende run kan zich richten op de invoer van Metis records die meerdere treffers in de thesaurus opleveren, waarbij een thesaurusrecord met overeenkomende leefjaren geselecteerd wordt om lokale en exemplaargebonden gegevens aan toe te voegen en in een vierde run worden records ingevoerd die alleen treffers met verschillende leefjaren opleveren.

Op deze wijze wordt het aantal Metis records dat nog verwerkt moet worden steeds kleiner en kan de matching telkens per run gericht op de nog overgebleven records worden afgesteld. De hier beschreven volgorde en het aantal uit te voeren runs kan in de praktijk, op basis van de opgedane ervaringen, worden aangepast.

Uiteindelijk blijft er een categorie records over waarbij het niet duidelijk is of er geen, één of meerdere treffers in de thesaurus aanwezig zijn. Deze records zullen een gelijkheidswaarde hebben die tussen de onder- en bovengrens zit, waarbij er geen mogelijkheden meer zijn om deze grenzen nog dichter bij elkaar te brengen. Deze categorie krijgt een bijzondere status, de zogenaamde B-status. Namen met een B-status hebben een link (ppn-verwijzing) naar het meest gelijkende record in de thesaurus. De B-status records dienen door deskundigen online bekeken te worden met WinIBW, de catalogiseerclient van het Pica systeem. Daarbij kan gebruik gemaakt worden van een aantal speciale functies zoals:

- het naast elkaar tonen op één scherm van het Metis record en het meest gelijkende thesaurusrecord
- het omzetten van de B-status in een normale status als de records verschillend zijn
- het hevelen van de lokale en exemplaargebonden velden naar het thesaurusrecord als de records hetzelfde zijn
- het snel kunnen verwijderen van overbodig geworden records met B-status.

Ten einde namen met een B-status efficiënt te kunnen verwerken zullen een aantal WinIBW scripts vervaardigd worden; zie Appendix 5: samenvoegen en splitsen van thesaurusrecords.

In de voorbereidingsfase zullen de namen die meerdere treffers opleveren niet direct als record met een B-status worden ingevoerd, maar zullen eerst lijsten worden gemaakt waarmee onderzocht worden of het matchen nog verder aangescherpt kan worden door, bijvoorbeeld, de index definitie aan een specifieke situatie aan te passen. De stapsgewijze aanpak is er op gericht om de onder- en bovengrenzen van de gelijkheidswaarde zo dicht mogelijk bij elkaar te brengen, zodat het aantal namen met een B-status zo klein mogelijk gehouden kan worden.

I Appendix: Metis export format

Voor de toelevering van onderzoekergegevens vanuit lokale (Metis) onderzoeksdatabanken worden twee export formats gehanteerd: een eenvoudige structuur gebaseerd op een vaste record en veld lengten en een URL gebaseerde structuur. De structuur met vaste lengten wordt zowel voor de initiële vulling als voor het (dagelijks) verwerken van lokaal aangebrachte updates gebruikt (zie Initiële vulling en Ophalen van Metis wijzigingen). De URL gebaseerde structuur wordt gebruikt in de 'MAAK DAI'' URL die wordt samengesteld als een DAI (PPN) wordt opgehaald uit de persoonsnamenthesaurus (zie Ophalen DAI).

De structuur met vaste lengten.

Record lengte: 275 posities:

1. Positie 001 t/m 005: onderzoekernummer
2. Positie 008 t/m 043: naam medewerker
3. Positie 044 t/m 059: voorletter(s)
4. Positie 060 t/m 080: voorvoegsel
5. Positie 081 t/m 096: titulatuur
6. Positie 097 t/m 097: voorkeur
7. Positie 099 t/m 099: geslacht
8. Positie 101 t/m 111: geboortedatum
9. Positie 112 t/m 113: code functie
10. Positie 115 t/m 165: functie
11. Positie 166 t/m 173: code organisatie
12. Positie 175 t/m 225: organisatie
13. Positie 226 t/m 235: begin aanstelling
14. Positie 237 t/m 246: eind aanstelling

URL- structuur

De URL die door Metis wordt opgestuurd naar de thesaurus om een DAI op te halen (zie Ophalen DAI) heeft de volgende opbouw:

	Attribuut	**Voorbeeld**
1	Adres OCLC PICA web pagina	http://develop.pica.nl:18080/
2	Usernummer/ wachtwoord	login/FORM/REQUEST?DB=1.1&USER_KEY=<usernummer> &PASSWORD=<wachtwoord>
3	Redirect	&REDIRECT=http%3A%2F%2Fdevelop.pica.nl%3A18080%2FDB %3D1.1
4	Zoekstring	%2FCMD%3FACT%3DSRCH%26IKT%3D1%26SRT%3DRLV%26REC%3D2 %26TRM%3Djansen%2c%20jan
5	Onderzoekernummer	%2Fp%5Fonderzoekernummer%3D20045
6	Naam mdewerker	%2Fp%5Fnaam%5Fmedewerker%3DObdam
7	Voorletter(s)	%2Fp%5Fvoorletter%3DPh.J.
8	Voorvoegsel	%2Fp%5Fvoorvoegsel%3D
9	Titulatuur	%2Fp%5Ftitulatuur%3DDr.

10	Voorkeur	%2Fp%5Fvoorkeur%3DJ
11	Geslacht	%2Fp%5Fgeslacht%3DV
12	Geboortedatum	%2Fp%5Fgeboortedatum%3D23-10-1953
13	Code functie	%2Fp%5Fcode%5Ffunctie%3D30
14	Functie	%2Fp%5Ffunctie%3DUniversitair docent
15	Code organisatie	%2Fp%5Fcode%5Forganisatie%5Fa%3D22020100
16	Organisatie	%2Fp%5Forganisatie%5Fa%3Dradiologie
17	Begin aanstelling	%2Fp%5Fbegin%5Faanstelling%3D01-01-2000
18	Einde aanstelling	%2Fp%5Feinde%5Faanstelling%3D31-10-2005

Voorbeeld:
http://develop.pica.nl:18080/login/FORM/REQUEST?DB=1.1&USER_KEY=3&PASSWORD=dai&REDIRECT=http%3A%2F%2Fdevelop.pica.nl%3A18080%2FDB%3D1.1%2FCMD%3FACT%3DSRCH%26IKT%3D1%26SRT%3DRLV%26REC%3D1%26TRM%3Djansen%2c%20jan%2Fp%5Fonderzoekernummer%3D20045 , etc.

Opmerkingen:

- Alle gegevens na de REDIRECT dienen gecodeerd (encoded) te worden.
- In een productie omgeving zal het adres van de Pica webpagina, het usernummer en het wachtwoord gewijzigd worden; usernummer en wachtwoord zullen per onderzoeksinstelling verschillend zijn.
- Mogelijk gebruik van volledige voornamen naast of in plaats van initialen moet na oplevering van een nieuwe Metis release nog nader worden uitgewerkt.
- Als tekenset wordt UTF8 gebruikt.

II Appendix: Pica export format

In het kader van het DAI project wordt een XML definitie opgesteld van het format dat is beschreven in hoofdstuk 4, uitgebreid met enkele velden afkomstig uit het thesaurusrecord, zoals het PPN (DAI), de volledige naam en, indien van toepassing het oude PPN – oude DAI.
De XML export van persoonsnamen met daaraan toegevoegde onderzoeksblokken is zowel in online als in offline omgeving mogelijk. De specificatie van de XML structuur bestaat uit twee delen: het XML-schema (XSD) en het bijbehorende XML document.

II.I XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
XML Schema for validating XML descriptions of the DAI export.
Version : 0.4
Creator : Hans Mugge
Date    : 03-04-2006
-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
    <!-- definition of simple elements -->
    <xs:element name="PPN_003at_797" type="xs:string"/>
    <xs:element name="EmployeeName_128Adollara_A10" type="xs:string"/>
    <xs:element name="Initials_128Adollard_A10" type="xs:string"/>
    <xs:element name="Prefix_128Adollarc_A10" type="xs:string"/>
    <xs:element name="Titles_128Adollare_A10" type="xs:string"/>
    <xs:element name="PreferredName_128A_A10" type="xs:string"/>
    <xs:element name="ResearchNumber_103M_A01" type="xs:string"/>
    <xs:element name="DateOfBirth_132Adollara_A30" type="xs:string"/>
    <xs:element name="Sex_132Adollarb_A30" type="xs:string"/>
    <xs:element name="tPersonalia_028A" type="xs:string"/>
    <xs:element name="tNameVariant_028at" type="xs:string"/>
    <xs:element name="FunctionName_232Bdollarb_B21" type="xs:string"/>
    <xs:element name="FunctionCode_232Bdollara_B21" type="xs:string"/>
    <xs:element name="OrganisationName_229Adollara_B10" type="xs:string"/>
    <xs:element name="OrganisationCode_229AdollarB_B10" type="xs:string"/>
    <xs:element name="StartDate_232Adollara_B20" type="xs:string"/>
    <xs:element name="EndDate_232Adollarb_B20" type="xs:string"/>
    <xs:element name="daiurl" type="inhoud"/>
    <!-- definition of complex elements -->
    <xs:complexType name="inhoud">
        <xs:sequence>
            <xs:element name="PersonalData">
                <xs:complexType>
                    <xs:sequence>
                        <xs:element ref="PPN_003at_797"/>
                        <xs:element name="Name">
                            <xs:complexType>
                                <xs:sequence>
                                    <xs:element ref="EmployeeName_128Adollara_A10"/>
                                    <xs:element ref="Initials_128Adollard_A10"/>
                                    <xs:element ref="Prefix_128Adollarc_A10" minOccurs="0"/>
                                    <xs:element ref="Titles_128Adollare_A10" minOccurs="0"/>
                                    <xs:element ref="PreferredName_128A_A10" minOccurs="0"/>
                                </xs:sequence>
                            </xs:complexType>
                        </xs:element>
                        <xs:element ref="ResearchNumber_103M_A01"/>
                        <xs:element ref="DateOfBirth_132Adollara_A30" minOccurs="0"/>
                        <xs:element ref="Sex_132Adollarb_A30" minOccurs="0"/>
                        <xs:element name="ThesaurusData">
                            <xs:complexType>
                                <xs:sequence>
                                    <xs:element ref="tPersonalia_028A" minOccurs="0"/>
                                    <xs:element ref="tNameVariant_028at" minOccurs="0"
maxOccurs="unbounded"/>
                                </xs:sequence>
                            </xs:complexType>
                        </xs:element>
                    </xs:sequence>
                </xs:complexType>
```

```
            </xs:element>
            <xs:element name="FunctionalData">
                <xs:complexType>
                    <xs:sequence>
                        <xs:element name="Function">
                            <xs:complexType>
                                <xs:sequence>
                                    <xs:element ref="FunctionName_232Bdollarb_B21" minOccurs="0"
maxOccurs="unbounded"/>
                                    <xs:element ref="FunctionCode_232Bdollara_B21" minOccurs="0"
maxOccurs="unbounded"/>
                                </xs:sequence>
                            </xs:complexType>
                        </xs:element>
                        <xs:element name="Organisation">
                            <xs:complexType>
                                <xs:sequence>
                                    <xs:element ref="OrganisationName_229Adollara_B10"
minOccurs="0" maxOccurs="unbounded"/>
                                    <xs:element ref="OrganisationCode_229AdollarB_B10"
minOccurs="0" maxOccurs="unbounded"/>
                                </xs:sequence>
                            </xs:complexType>
                        </xs:element>
                        <xs:element name="Appointment">
                            <xs:complexType>
                                <xs:sequence>
                                    <xs:element ref="StartDate_232Adollara_B20" minOccurs="0"
maxOccurs="unbounded"/>
                                    <xs:element ref="EndDate_232Adollarb_B20" minOccurs="0"
maxOccurs="unbounded"/>
                                </xs:sequence>
                            </xs:complexType>
                        </xs:element>
                    </xs:sequence>
                </xs:complexType>
            </xs:element>
        </xs:sequence>
    </xs:complexType>
</xs:schema>
```

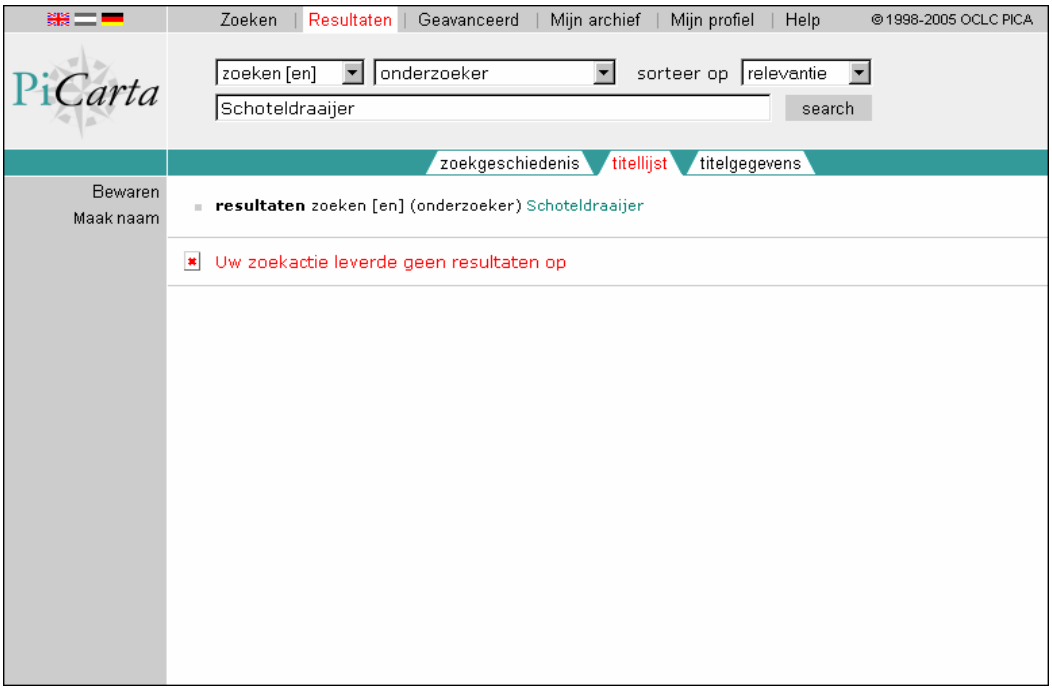
II.II XML document

```
  <?xml version="1.0" encoding="utf-8" ?>
<catalog banner="NTA">
<record>
            <seqnr>1</seqnr>
            <sortkey>067745560 </sortkey>
            <main>067575560</main>
            <PersonalData>
                        <PPN_003at_797>067745560</PPN_003at_797>
                        <PPN_103Z_A99>044596941<PPN_103Z_A99>
                        <Name>

<EmployeeName_128Adollara_A10>Schijndel</EmployeeName_128Adollara_A10>
                                    <Initials_128Adollard_A10>K.J.</Initials_128Adollard_A10>
                                    <Prefix_128Adollarc_A10>van</Prefix_128Adollarc_A10>
                                    <Titles_128Adollare_A10>Prof.Dr.</Titles_128Adollare_A10>
                                    <PreferredName_128A_A10>J.</PreferredName_128A_A10>
                        </Name>
                        <ResearchNumber_103M_A01>06817</ResearchNumber_103M_A01>
                        <DateOfBirth_132Adollara_A30>15-11-1955</DateOfBirth_132Adollara_A30>
                        <Sex_132Adollarb_A30>M</Sex_132Adollarb_A30>
                        <ThesaurusData>
                                    <tPersonalia_028A>Karel Jan van Schijndel, 15-11-1955;
professor</tPersonalia_028A>
```

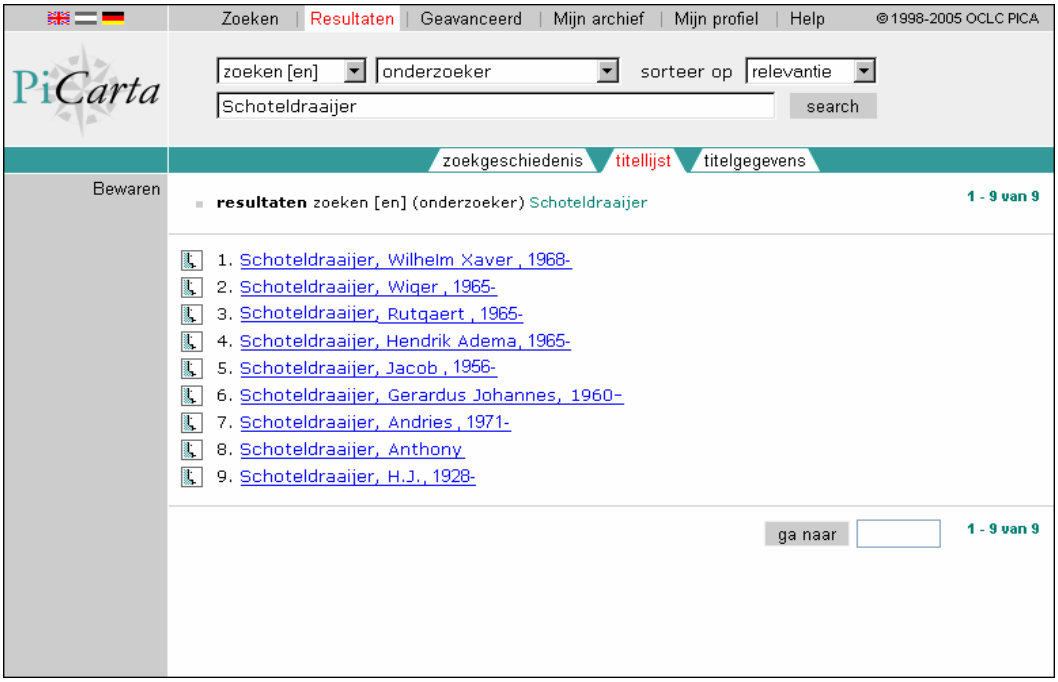
```
                                    <tNameVariant_028at>Carel Jan van
Schijndel</tNameVariant_028at>
                        </ThesaurusData>
            </PersonalData>
            <FunctionalData>
                        <Function>
                                    <FunctionName_232Bdollarb_B21>Universitair
hoofddocent</FunctionName_232Bdollarb_B21>

<FunctionCode_232Bdollara_B21>20</FunctionCode_232Bdollara_B21>
                        </Function>
                        <Organisation>
                                    <OrganisationName_229Adollara_B10>Huisartsgeneeskunde-
vakgroep</OrganisationName_229Adollara_B10>

<OrganisationCode_229Adollarb_B10>22000200</OrganisationCode_229Adollarb_B10>
                        </Organisation>
                        <Appointment>
                                    <StartDate_232Adollara_B20>01-10-
1991</StartDate_232Adollara_B20>
                                    <EndDate_232Adollarb_B20>01-10-
2001</EndDate_232Adollarb_B20>
                        </Appointment>
                        <Function>

<FunctionName_232Bdollarb_B21>Onderzoeker</FunctionName_232Bdollarb_B21>

<FunctionCode_232Bdollara_B21>27</FunctionCode_232Bdollara_B21>
                        </Function>
                        <Organisation>

<OrganisationName_229Adollara_B10>Oogheelkunde</OrganisationName_229Adollara_B10>

<OrganisationCode_229Adollarb_B10>22700900</OrganisationCode_229Adollarb_B10>
                        </Organisation>
                        <Appointment>
                                    <StartDate_232Adollara_B20>01-10-
2001</StartDate_232Adollara_B20>
                        </Appointment>
            </FunctionalData>
</record>
</catalog>
```

III Appendix: Thesaurus templates

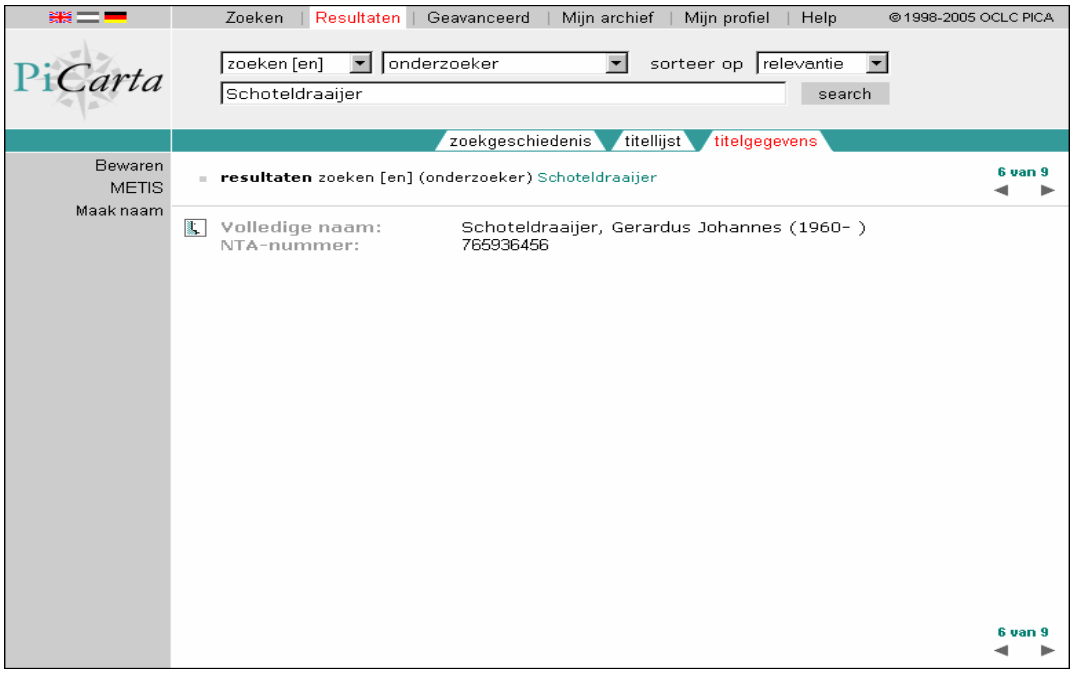
Deze Appendix bevat voorbeelden van drie templates die getoond worden bij het zoeken in de Nederlandse Thesaurus voor Auteursnamen. Het template dat gebruikt worden voor het invoeren van Metis gegevens in the thesaurus en het template dat getoond wordt nadat de in te voeren gegevens zijn geaccepteerd, staan in Ophalen DAI.



Template Geen naam gevonden



Template meerdere namen gevonden



Template één naam gevonden of gekozen uit namenoverzicht

IV Appendix: Kwaliteitseisen

Het DAI project is een pilot project dat bij welslagen als voorbeeld zal dienen voor gelijksoortige projecten bij de overige Nederlandse onderzoeksinstellingen. Daarom is het van belang om de kwaliteitseisen vast te stellen waaraan de gegevens, de procedures en de documentatie dienen te voldoen, zodat vervolgprojecten optimaal gebruik kunnen maken van de resultaten van het pilot project.

IV.I Gegevens

De geformuleerde eisen hebben betrekking op de gegevens die worden aangeleverd vanuit de onderzoeksbestanden; de kwaliteit van de gegevens in de persoonsnamenthesaurus wordt geaccepteerd zoals deze is. Voor de initiële vulling en voor de batch verwerking van in onderzoeksbestanden aangebrachte wijzigingen gelden de volgende kwaliteitseisen:

1. De naamsgegevens dienen zo volledig mogelijk te zijn:
   - Zoveel als mogelijk is de volledige voornamen in plaats van of naast de initialen
   - Achternaam en voornaam zijn verplichte subvelden; namen waarin één van beide ontbreekt, worden niet verwerkt
   - Geboortejaren zijn dringend gewenst; bij matchen in thesaurus te gebruiken ter onderscheiding van verschillende personen; de uit onderzoeksbestanden overgenomen geboortejaren kunnen desgewenst in presentaties niet getoond worden.
   - Geslacht is een verplicht veld; bij namen zonder dit veld wordt als inhoud ‘o’ (onbekend) gegenereerd.
   - De Voorkeursindicatie is een verplicht veld bij de initiële vulling; velden zonder deze indicatie worden niet verwerkt.

2. De onderzoeksgegevens dienen zo volledig mogelijk te zijn:
   - Naam van de organisatie, Omschrijving van de functie en Begindatum van de aanstelling zijn verplichte velden; namen zonder deze velden worden niet verwerkt
   - Code van de organisatie en code van de functie zijn gewenst, maar niet verplicht

3. Specifieke eisen voor (dagelijkse) batch verwerking:
   - DAI (PPN) is een verplicht veld; namen zonder DAI worden in de dagelijkse batch verwerking niet behandeld.
   - Namen zonder Voorkeursindicatie worden als naamsvariant toegevoegd; indien al één of meer naamsvarianten in het lokale blok aanwezig zijn, worden identieke naamsvarianten ontdubbeld.

IV.II Procedures

De geformuleerde eisen hebben betrekking op alle procedures die in het pilot DAI project en in vervolgprojecten voor zullen komen.

1. Aanleveren van Naams- en onderzoeksgegevens voor de initiële vulling
   - De gegevens worden beschikbaar gesteld op een FTP site waar Pica ze kan ophalen; in geval van wachtwoordprotectie, wordt het benodigde wachtwoord ter beschikking gesteld.
   - De file en datastructuur dienen stabiel en conform specificatie te zijn; incidentele wijzigingen dienen tijdig te worden doorgegeven aan de Servicedesk van OCLC PICA.
   - Voor elke onderzoeker wordt één naam met voorkeursindicatie aangeleverd. Bijbehorende Naamsvarianten kunnen in tweede instantie worden aangeleverd, nadat de DAI’s in de onderzoeksdatabank zijn opgenomen.

2. Conversie van de gegevens voor de initiële vulling
   - Bij de conversie mogen geen gegevens verloren gaan.

3. Initiële vulling
   - De persoonsnamenthesaurus dient te beschikken over een voor aanwezigheidscontrole geoptimaliseerde index
   - Mogelijk dubbel ingevoerde namen dienen zodanig gekenmerkt te worden, dat regulier gebruik van de thesaurus zo weinig mogelijk wordt gehinderd
   - Voor mogelijk dubbel ingevoerde namen dienen efficiënte online opschoningvoorzieningen ter beschikking gesteld te worden:
     - Speciale index om deze namen snel te kunnen zoeken
     - Functies voor
       i. Het vergelijken van mogelijk dubbele namen
       ii. Het hevelen van Metis naamsgegevens en onderzoeksgegevens naar de voorkeursnaam
       iii. Het verwijderen van dubbel ingevoerde namen

4. DAI export na de initiële vulling
   - Complete export van namen met aangehechte Metis namen en onderzoeksblokken, inclusief DAI (PPN) en Metis nummers dienen via FTP beschikbaar gesteld te worden aan de lokale onderzoeksdatabank na elke fase van het matching proces
   - In Metis dient voor DAI een geïndexeerd veld gedefinieerd te worden

5. Aanmaken / ophalen DAI
   - Onnodig tikwerk dient te worden vermeden:
     - Automatische login
     - Genereren van zoeksleutels uit gevonden namen
     - Meegeven van Naams- en onderzoeksgegevens in URL
     - Invullen van Naams- en onderzoeksgegevens in de web templates van de persoonsnamenthesaurus
   - Aanmaken en ophalen van DAI's moet mogelijk zijn voor medewerkers zonder bibliotheekopleiding:
     - Duidelijke help bij zoeken, invoer / wijzigen van gegevens en bij het heen en weer schakelen tussen Metis en thesaurus

6. Verwerken van lokale wijzigingen
   - De gegevens die in Metis gewijzigd worden dienen dagelijks automatisch verzameld, geconverteerd en op een voor Pica toegankelijke FTP site geplaatst te worden; per onderzoeker worden alle gegevens (alle naamsvarianten en alle onderzoeksgegevens aangeboden.
   - De te verwerken gegevens worden geïdentificeerd met de bijbehorende DAI; gegevens zonder DAI worden niet verwerkt
   - Een geautomatiseerd proces bij Pica controleert dagelijks de daartoe bestemde FTP sites, converteert de gevonden gegevens naar het thesaurus format en verwerkt de gegevens:
     - Namen zonder DAI worden opgeslagen in een error log
     - Bij namen met een DAI worden in de thesaurus voor de betreffende bibliotheek / universiteit alle lokale (Metis namen) en exemplaarvelden (onderzoeksgegevens) in de thesaurus vervangen door de nieuwe velden, met voorzover mogelijk, handhaving van de reeds toegekende productienummers (epn's).

IV.III Documentatie

De kwaliteitseisen voor documentatie hebben betrekking op gebruikersdocumentatie, de beschrijving van de te volgen procedures en de beschrijvingen van de toegepaste datastructuren.

1. Gebruikersdocumentatie
   - o Gebruikersdocumentatie is in het Nederlands
   - o Gebruikersdocumentatie is actueel
   - o Voor de web interface wordt gebruikersdocumentatie uitsluitend geleverd als online help
   - o Het taalgebruik in de online help is gericht op de gebruikersgroep: medewerkers zonder bibliotheekervaring
   - o Voor bibliotheekmedewerkers worden een specialistische handleiding opgesteld voor het opschonen van mogelijk dubbel ingevoerde namen

2. Procedurebeschrijvingen
   - o De procedurebeschrijvingen zijn in het Nederlands
   - o Het taalgebruik van de procedurebeschrijvingen is gericht op de gebruikersgroep: beheerders van onderzoekdatabanken
   - o Procedurebeschrijvingen zijn benodigd voor:
     - Aanleveren van initiële vullingen in de voorgeschreven datastructuur
     - Aanleveren van dagelijkse wijzigingen in de voorgeschreven structuur
     - Behandeling van fouten

3. Documentatie van datastructuren
   - o Documentatie van Metis datastructuren en het thesaurus format is in het Nederlands
   - o Documentatie van het thesaurus format is een onderdeel van de algemene thesaurus documentatie en toegankelijk via de OCLC Pica website

IV.IV Systeemeisen

De volgende kwaliteitseisen worden aan de technische infrastructuur gesteld:

1. Ondersteunde browsers

In beginsel kunnen alle browsers gebruikt worden die CSS (Cascading Style Sheets) ondersteunen:

- o Internet Explorer vanaf v5.0
- o Netscape vanaf v6.1
- o Mozilla vanaf v1.0
- o Firefox vanaf v1.0
- o Opera vanaf v6.0
- o Safari vanaf v1.0

2. Catalogiseer client WinIBW
   - o WinIBW 2.x met gebruik van Intermarc tekenset (Pica tekenset)
   - o WinIBW 3.x met gebruik van UTF8 tekenset

3. Tekenset
   - o Voor web interface: UTF8
   - o Voor WInIBW: Intermarc (WinIBW 2.x) of UTF8 (WinIBW 3.x)

V Appendix: Samenvoegen en splitsen van thesaurusrecords

Bij de initiële vulling van Naams- en onderzoekgegevens worden onderzoeksblokken aan namen in de thesaurus toegevoegd en worden nieuwe namen met onderzoeksblokken aangemaakt. De laatste categorie bestaat uit twee onderdelen: namen die bij de aanwezigheidscontrole niet zijn gevonden en namen waarbij de software geen besluit kon nemen over het al dan niet aanwezig zijn. Namen die tot de laatste groep behoren zijn ingevoerd met een speciale status, de B-status en hebben een extra veld waarin het PPN is opgenomen van de naam die de meeste gelijkenis vertoont met de ingevoerde naam. De namen met een B-status dienen online geanalyseerd en opgelost te worden. Voor het samenvoegen en splitsen van namen is specifieke vakkennis noodzakelijk die gewoonlijk alleen in de bibliotheek van een onderzoeksinstelling beschikbaar is. Bibliotheekmedewerkers gebruiken voor hun dagelijks werkzaamheden met het Pica systeem WinIBW, een windows client uitgerust met specifieke functies voor het bewerken van titelbeschrijvingen en thesaurusrecords. WinIBW zal ook voor het samenvoegen en splitsen van namen van onderzoekers gebruikt worden. Hiervoor zullen een aantal WinIBW scripts vervaardigd worden die hieronder worden beschreven.

Samenvoegen en splitsen van thesaurusrecords kan ook nodig zijn bij namen die geen B-status hebben. Die gevallen lijken meer op de opschoonacties die met name door medewerkers van de Koninklijke Bibliotheek worden uitgevoerd. De voor deze acties benodigde handelingen blijven hier vooralsnog buiten beschouwing. Indien nodig kunnen hiervoor later nog één of meer scripts vervaardigd worden.

Namen met een B-status hebben een aparte index die is opgebouwd uit de elementen van het Pica+ veld 038L (zie Velden op gemeenschappelijk niveau). Deze index zorgt er voor dat elke onderzoeksinstelling in één zoekactie alle 'eigen' B-records kan verzamelen, die vervolgens met WinIBW bekeken kunnen worden[2]. Deze online controle dient uit te wijzen of een naam met B-status wel of niet dubbel is. Voor beide situaties zal een script beschikbaar zijn dat er voor zorgt dat de gewenste situatie met zo weinig mogelijke manuele handelingen gerealiseerd kan worden.

Als geconstateerd wordt dat een naam dubbel is, moeten twee thesaurusrecords worden samengevoegd tot één. Daarbij geldt de volgende basisregel:

- Indien één van de twee namen gekoppeld is aan titelbeschrijvingen in het GGC, dient deze naam gehandhaafd te worden en dienen gegevens van de andere naam te worden overgeheveld .
- Als geen van de dubbele namen aan titelbeschrijvingen is gekoppeld, kan men zelf bepalen welke naam de voorkeur moet krijgen.

In beide gevallen moeten de onderzoeksblokken naar de voorkeursnaam worden overgeheveld, waarna de andere naam verwijderd kan worden.

Voor het hevelen van onderzoeksblokken naar de voorkeursnaam wordt een script ter beschikking gesteld dat de volgende functionaliteit zal hebben:

- Leest het PPN in veld 038L (PPN van voorkeursnaam),
- Leest het PPN van de naam met B-status
- Plakt het PPN van de naam met B-status in veld 103Z

[2] De functie die voor B-titels wordt gebruikt, waarbij men op basis van 038L twee titels naast elkaar op het scherm kan bekijken en waar men door OK of NOK aan te klikken kan aangeven of de titels al dan niet gelijk zijn, wordt voor namen niet geschikt geacht, omdat bij titels de verdere afhandeling via complexe batch procedures plaats vindt, die voor namen niet nodig zijn.

- Gebruikt het PPN van de voorkeursnaam in het hevelcommando dat ervoor zorgt dat alle onderzoeksblokken van de betreffende onderzoeksinstelling aan de voorkeursnaam worden gekoppeld[3]
- Genereert een verwijdercommando (inclusief de afgedwongen herbevestiging) voor de naam met de B-status.
- Keert terug naar de volgende regel in het overzicht van namen met B-status, zodat er met de volgende naam kan worden doorgegaan.

In de toekomst kan het voorkomen dat er dubbele namen gevonden worden met onderzoeksblokken van verschillende onderzoeksorganisaties, dus van verschillende ILN's. In die gevallen kan men alleen de 'eigen' onderzoeksblokken hevelen naar de voorkeursnaam. Indien een naam verwijderd moet worden waaraan onderzoeksblokken zijn toegevoegd van een ander ILN, dient de betreffende onderzoeksinstelling daarvan op de hoogte te worden gebracht. Zonodig kan hiervoor later nog een script vervaardigd worden.

Als een naam met B-status niet dubbel is, dient de B-status ongedaan te worden gemaakt. Hiervoor wordt een script gemaakt dat de volgende functionaliteit zal hebben:
- Activeert de WinIBW editor voor de betreffende naam.
- Verandert de B in veld 005 in een 'x' waardoor de B-status ongedaan is gemaakt.
- Verwijdert veld 038L (796).
- Stuurt de gemuteerde naam naar de server.
- Keert terug naar de volgende regel in het overzicht van namen met B-status, zodat er met de volgende naam kan worden doorgegaan.

[3] Hevelen is in de huidige versie van het systeem alleen bij titels mogelijk; in afwachting van de volgende versie zal het hevelen in het script tijdelijk worden vervangen door een Delete bij de B-naam, gevolgd door een Insert bij de voorkeursnaam.