Digital Author Identification (DAI)

Anneloes Degenaar a.l.degenaar@rug.nl

Een persoonsnamenthesaurus

In het Digital Author Identification (DAI)-project is gewerkt aan de totstandkoming van een auteursnamenthesaurus, de Nederlandse Thesaurus van Auteursnamen (NTA), van alle aan Nederlandse universiteiten werkende auteurs. Projectmanager Anneloes Degenaar geeft antwoord op vragen van Pictogram.

Waarom is een auteursidentificatie nodig?

Het doel van het DAI-project is om elke auteur, met een aanstelling of een andere relevante band bij een Nederlandse universiteit of onderzoeksinstituut, een uniek landelijk nummer toe te kennen (DAI-nummer):

a. De Nederlandse Thesaurus van Auteursnamen voorziet in een behoefte die bestaat aan een betere vindbaarheid van publicaties;
b. Het nummer dient om publicaties van eenzelfde auteur beter bij elkaar te kunnen vinden;
c. Het nummer dient om auteurs met dezelfde naam van elkaar te kunnen onderscheiden;
d. Het systeem sluit aan bij een praktijk die in bibliotheekcatalogi al jaren in gebruik is;
e. Het is een werknummer en het kan - in voorkomende gevallen - wijzigen;
f. Het gaat niet om een nieuw op te zetten persoonsregistratie.

DAI helpt de wetenschapper in zijn wetenschappelijke communicatie. Door de toekenning van een uniek nummer is het in de toekomst mogelijk om alle publicaties van een onderzoeker, ongeacht of hij bij verschillende universiteiten heeft gewerkt, eenvoudig bij elkaar te vinden.

Welke instellingen hebben aan dit project meegedaan?

De participerende instellingen aan dit project waren de Universiteitsbibliotheek van de Rijksuniversiteit Groningen (penvoerder), OCLC PICA en het Universitair Centrum Informatievoorziening van de Radboud Universiteit (UCI).

In het DAI-project is ervoor gekozen DAI eerst als een pilot bij de Rijksuniversiteit Groningen in productie te brengen. De kennis, de producten en ervaringen die bij de Rijksuniversiteit Groningen zijn opgedaan zijn essentieel voor de verdere landelijke implementatie. De uitrol van DAI naar andere universiteiten wordt gerealiseerd via een vervolgproject: DAI Uitrol. Het gaat om de volgende instellingen:

- Universiteit Twente
- Universiteit van Amsterdam
- Erasmus Universiteit Rotterdam
- Radboud Universiteit Nijmegen
- Technische Universiteit Delft
- Technische Universiteit Eindhoven
- Universiteit Leiden
- Universiteit Maastricht
- Universiteit Utrecht
- Universiteit van Tilburg
- Vrije Universiteit Amsterdam
- Wageningen Universiteit & ResearchCentrum
- Centrum voor Wiskunde en Informatica (CWI) (Niet-METIS gebruiker)

Het project DAI Uitrol heeft als doel DAI operationeel te maken voor alle andere deelnemende instellingen van het programma Digital Academic Repositories (DARE), die het onderzoeksregistratiesysteem METIS gebruiken. Hierdoor zal de waarde van de Nederlandse Thesaurus van Auteursnamen toenemen en wordt de integratie van systemen en repositories verbeterd. Hiermee wordt een basis geboden van waaruit de kwaliteit van het zoeken naar elektronische publicaties sterk kan worden verbeterd. Met behulp van het unieke DAI-nummer wordt de mogelijkheid geboden om een zo volledig en nauwkeurig mogelijke selectie te maken van de publicaties van een

auteur, verbonden aan een Nederlandse universiteit of onderzoeksinstelling.
De verwachting is dat eind 2006 de Nederlandse Thesaurus van Auteursnamen door alle universiteiten is gevuld en daardoor een belangrijk identificatiesysteem is geworden.

Welke auteurs zitten er in de database?

In het DAI-project is ervoor gekozen om alle actieve wetenschappers te thesaureren. Dat betekent dat alle onderzoekers (auteurs) met een aanstelling of een andere relevante band bij een Nederlandse universiteit of onderzoeksinstituut, die publiceren, een uniek landelijk nummer krijgen toegekend.

Bestond er al zoiets als een thesaurus van auteursnamen?

De Nederlandse Thesaurus van Auteursnamen is gebaseerd op de persoonsnamenthesaurus van OCLC PICA die door de bibliotheken wordt aangevuld en bijgehouden. Alle auteurs die hierin worden opgenomen, krijgen automatisch een uniek nummer toegekend (een zogenoemd Pica Productie Nummer, PPN). Dit nummer wordt gebruikt als DAI-nummer en wordt opgenomen in de verschillende lokale onderzoeksregistratiesystemen (METIS). De persoonsnamenthesaurus is een onderdeel van het GGC, het Gemeenschappelijk Geautomatiseerde Catalogiseersysteem van OCLC PICA. Voor dit project is de persoonsnamenthesaurus ook als een zelfstandige module beschikbaar gesteld. Dit betekent dat er blijvend door de bibliotheken kan worden ingevoerd, maar ook dat er een aparte zoek- en invoermodule is gemaakt waarmee vanuit de verschillende onderzoekssystemen in de Nederlandse Thesaurus van Auteursnamengezocht kan worden.

Hoe zijn alle auteursgegevens in de database ingevoerd?

Het UCI heeft een script ontwikkeld om gegevens (de initiële vulling) als batch vanuit METIS uit te schrijven zodat deze kunnen worden ingelezen in de Nederlandse Thesaurus van Auteursnamenbij OCLC PICA.
Vervolgens worden de gegevens bij OCLC PICA ingelezen in de NTA, waarna het proces van "matchen & mergen" plaatsvindt. Dit werkt als volgt:
Voor elke persoon die wordt aangeleverd, wordt met behulp van de criteria achternaam, voornaam, initialen en geboortejaar automatisch gecontroleerd of deze al in de Nederlandse Thesaurus van Auteursnamenaanwezig is.
Per persoon kan de controle drie uitkomsten hebben:

- De persoon is aanwezig in de NTA, de onderzoeksgegevens worden aan deze persoon toegevoegd.
- De persoon is niet aanwezig in de NTA, er wordt een nieuwe persoon ingevoerd, de onderzoeksgegevens worden aan deze persoon toegevoegd.
- Het controlemechanisme denkt dat de persoon in de NTA aanwezig is, maar weet het niet zeker. Het controlemechanisme doet in dit geval per persoon een voorstel.

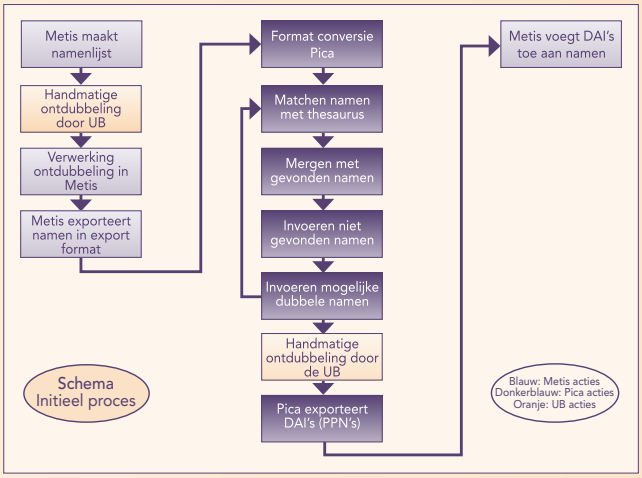
Alle personen uit de laatste categorie kunnen worden geselecteerd, waarna vervolgens (door bibliotheekmedewerkers) handmatig wordt gecontroleerd of het inderdaad de juiste persoon is die het controlemechanisme voorstelt.
Voor de nieuwe auteurs die in METIS worden ingevoerd, na de initiële vulling, kan in METIS een DAI-nummer worden opgevraagd vanuit de NTA.

Waren er problemen bij de matching?

Gebleken is dat het voor een nauwkeurige matching met de Nederlandse Thesaurus van Auteurs van belang is dat er zoveel mogelijk gegevens per auteur beschikbaar zijn.
Omdat METIS het bronbestand is voor de vulling van de NTA is het van belang dat METIS goed is geïmplementeerd. In dit project is gebleken dat het bronbestand van METIS RUG nog onvoldoende data bevatte om te kunnen gebruiken voor het DAI-project. Daarom is ervoor gekozen de data in METIS RUG te verrijken met data uit het personeelsinformatiesysteem. Dit bleek echter niet eenvoudig te zijn, waardoor het vrij lang heeft geduurd voor er een goed gevuld bestand uit METIS RUG kon worden opgeleverd voor het DAI-project.
Voor het vervolgtraject DAI Uitrol is het dus van belang dat andere universiteiten de data in METIS goed op orde hebben. Dit betekent dat er in elk geval een naam, voorletters, voorvoegsel, titel, geslacht, geboortejaar, functiebenaming en organisatienaam is opgenomen. Veel velden in METIS zijn niet verplicht om in te vullen, daarom ontbreken er ook vaak bepaalde gegevens.

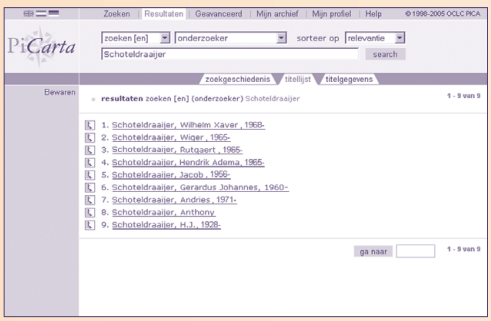
Kun je iets vertellen over de achterliggende techniek?

In het DAI-project is gekozen voor een gecontroleerde uitlevering van identificatienummers uit



In bovenstaand overzicht wordt getoond hoe de initiële vulling van de NTA met METIS-gegevens in zijn werk gaat.

Onderstaand een voorbeeld van de schermen die getoond worden bij het zoeken in de Nederlandse Thesaurus voor Auteursnamen.



de Nederlandse Thesaurus van Auteursnamen. Hierdoor verkrijg je unieke identificatienummers en wordt de kwaliteit van de database gewaarborgd. De uitlevering van NTA-nummers wordt in twee fasen gedaan.

De eerste fase is de initiële vulling, waarbij de gegevens van alle auteurs, die met de universiteit of onderzoeksinstelling een werkverband hebben, uit het onderzoeksregistratiesysteem van de universiteit worden binnengehaald in de Nederlandse Thesaurus van Auteursnamen en vervolgens een uniek nummer toegewezen krijgen.

Bij de initiële vulling wordt gecontroleerd of de auteur al aanwezig is in de database van de NTA of niet. Dit gebeurt via een proces van matching en merging. Bij twijfelgevallen wordt handmatig gecontroleerd of de auteur al in de database voorkomt of niet. De informatie over de auteurs met daarbij het DAI-nummer wordt vervolgens teruggestuurd naar de database van onderzoekers, bijvoorbeeld het onderzoeksregistratiesysteem METIS, waar deze informatie wordt ingelezen.

Na de initiële vulling worden de auteursgegevens primair in de database van de universiteit bijgehouden. Bij de Rijksuniversiteit Groningen is dat METIS, dat als bronsysteem wordt gezien. De controle over welke auteurs een DAI-nummer krijgen toegekend ligt bij de lokale invoerders. Hiervoor is gekozen, omdat lokaal de gegevens het best beschikbaar zijn. Periodiek zullen wijzigingen worden ge-upload in de Nederlandse Thesaurus van Auteursnamen.

De tweede fase bestaat uit de online workflow. Bij nieuwe onderzoekers wordt als zij ingevoerd worden in de database van de universiteit direct een DAI-nummer opgevraagd uit de Nederlandse Thesaurus van Auteursnamen. Hierbij wordt eerst een zoekactie in de NTA gedaan om te onderzoeken of de auteur zich al in de database bevindt. Mocht de auteur niet gevonden worden, dan wordt er een nieuw DAI-nummer toegekend.

Is er ook internationale commitment?

Bij de totstandkoming van DAI is ook rekening gehouden met internationale ontwikkelingen. Hierbij is onder meer gekeken naar de "outline" voor een ISO International Standard Party Identifier Code (ISPI). DAI en ISPI zijn compatibel. Als ISPI in 2008 ingevoerd gaat worden is er wel een project nodig om DAI en ISPI aan elkaar te koppelen.

Zijn er ook privacy-problemen?

De NTA thesaurus bevat mogelijk privacygevoelige gegevens over een auteur, zoals de geboortedatum. Deze gegevens worden alleen gebruikt om het proces van het toekennen van een nummer aan een naam te ondersteunen; ze worden niet naar buiten gebracht en kunnen niet ingezien worden door anderen dan degenen die professioneel bezig zijn met de registratie van wetenschappelijke output. Er wordt zeer zorgvuldig met deze gegevens omgegaan. De NTA-thesaurus bewaart extra informatie over een auteur, maar deze wordt nooit naar buiten gebracht. Daarbij geldt dat de METIS-beheerders alleen toegang hebben tot de volledige set gegevens van auteurs met een aanstelling bij de eigen instelling. Een METIS-beheerder van Utrecht kan niet de geboortedata van auteurs met een aanstelling bij de UvA inzien. Over het algemeen zijn auteurs gebaat bij vindbaarheid en dus ook bij een DAI-nummer.

Wat zijn de uiteindelijke resultaten? Wat moet er nog gebeuren in de toekomst?

Het DAI-project is succesvol afgerond. Er is een werkend systeem opgeleverd voor het toekennen van een landelijk uniek nummer aan auteurs en het zoeken van auteurs met behulp van dit landelijk nummer. Dit systeem is als pilot bij de RUG in productie gebracht, waarbij de in METIS ingevoerde auteurs voorzien zijn van een DAI-nummer.

Ook zijn de werkprocessen uitvoerig beschreven en is er voldoende technische informatie beschikbaar via het digitaal handboek op de website. Nu deze pilot is afgerond zal het komende najaar DAI operationeel worden voor alle andere deelnemende instellingen van het programma Digital Academic Repositories (DARE), die METIS gebruiken. De integratie van systemen die gebruikt worden voor wetenschappelijke informatievoorziening zal hierdoor bevorderd worden. Door DAI is er straks één gezamenlijke ingang op onderzoeksinformatie, waarbij de invalshoek niet disciplinegericht maar auteursgericht is.

Meer informatie over DAI en DAI Uitrol vindt u op:
dai.weblog.ub.rug.nl/
dai-uitrol.ub.rug.nl/