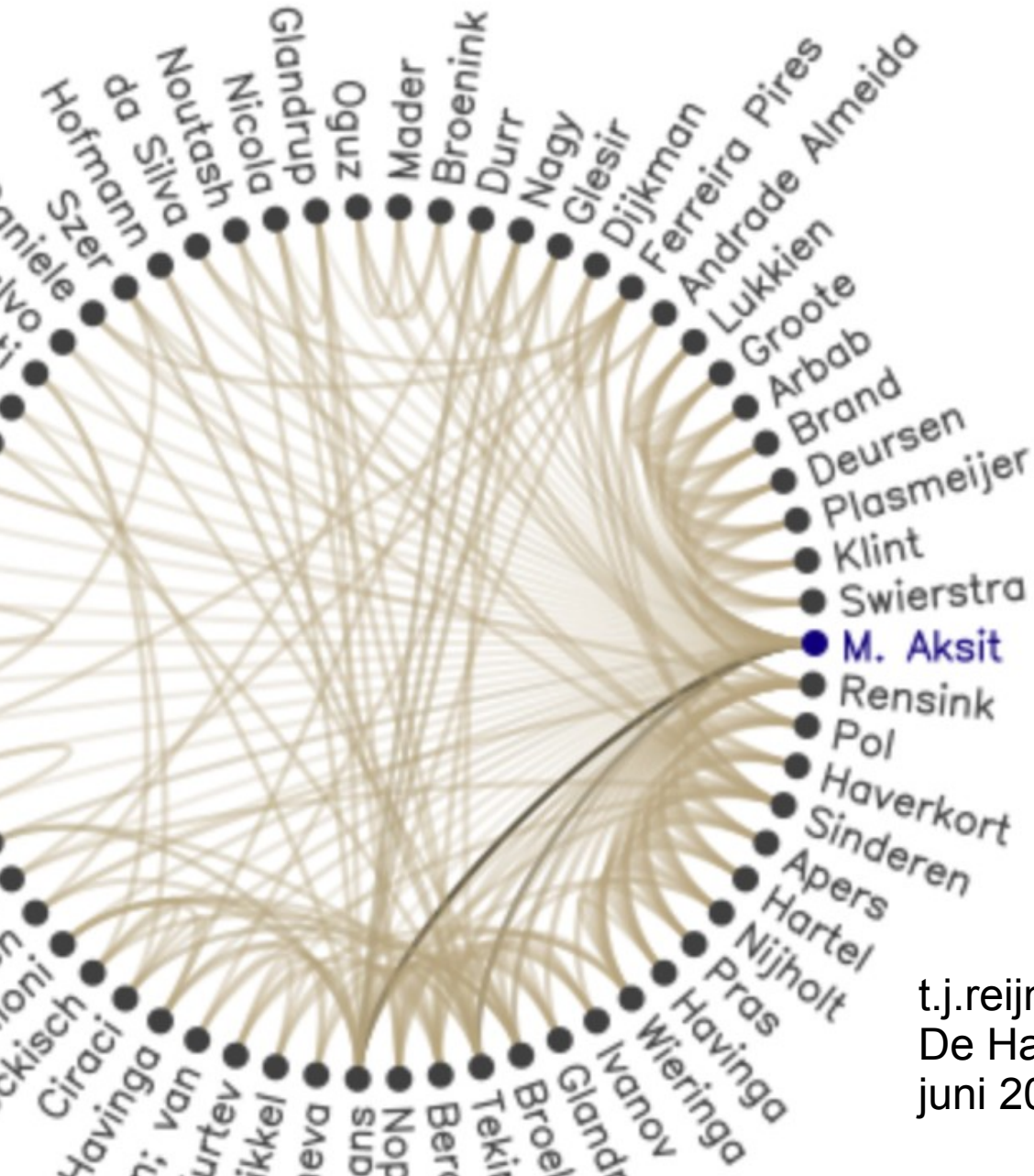
KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

de meerwaarde van de Digital Author Identification



t.j.reijnhoudt, 10102515
De Haagse Hogeschool
juni 2011

Voorwoord

Dit afstudeerverslag is geschreven in het kader van mijn afstuderen aan de Haagse Hogeschool, sector ICT & Media, richting Informatica.

Het afstudeerverslag is bedoeld voor het beschrijven en verantwoorden van mijn afstudeerperiode, zodat de examinatoren zich een beeld kunnen vormen van het verloop van het afstudeertraject.

Op deze plek wil ik ook graag mijn bedrijfsmentor Chris Baars bedanken voor zijn enthousiasme en betrokkenheid.

T.J. Reijnhoudt

Amsterdam, 1 juni 2011

Inhoudsopgave

1 Inleiding

Sinds 2007 ben ik werkzaam bij de afdeling Onderzoek Informatie van de KNAW als wetenschappelijk programmeur. Het grootste deel van de dagelijkse activiteiten van de programmeurs van de afdeling bestaat uit het onderhouden en vernieuwen van NARCIS. NARCIS staat voor National Academic Research and Collaborations Information System. De openbaar toegankelijke interface is bereikbaar op www.narcis.nl.

Als deeltijdstudent ligt het voor de hand om de eigen werkomgeving als afstudeerbedrijf te gebruiken. In overleg met de begeleider examinator Tim Cocx van de Haagse Hogeschool heb ik een afstudeeropdracht geformuleerd die voldoet aan de eisen en wensen van het afstuderen binnen de Academie voor ICT en Media. Omdat ik parttime medewerker ben aan de Haagse Hogeschool is er een externe expert examinator aangetrokken, dhr G.J. Timmerman van Info Support.

In overleg met de bedrijfsmentor Chris Baars is ervoor gezorgd dat de opdracht past binnen de doelstelling van de afdeling: het etaleren van Nederlands publiek-gefinancierd onderzoek, onderzoekers en instituten.

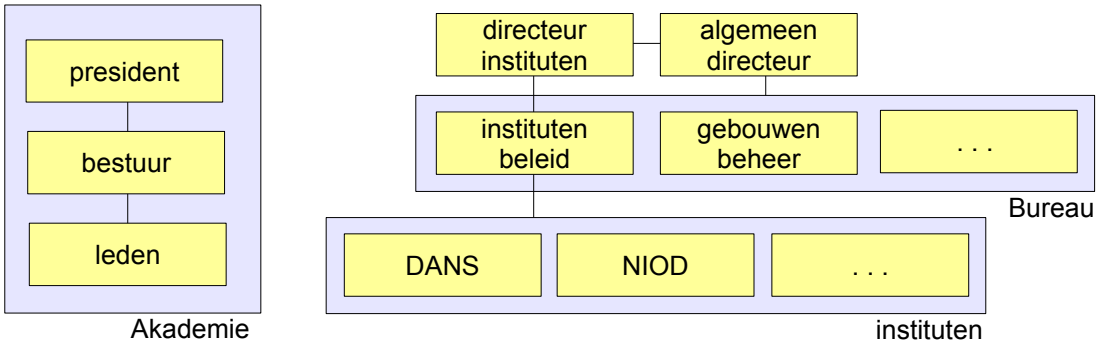
Tijdens het afstudeertraject is de afdeling Onderzoek Informatica onderdeel geworden van DANS[1], een instituut van de KNAW en NWO. Op dat moment is de adjunct directeur dr. Henk Harmsen de opdrachtgever geworden.

1 http://dans.knaw.nl

2 KNAW, de afdeling Onderzoek Informatie

2.1 De KNAW

De Koninklijke Nederlandse Akademie van Wetenschappen (KNAW) heeft als doelstelling de kwaliteit en de belangen van de wetenschap te bevorderen. Zij zet zich in voor een optimale bijdrage van de Nederlandse wetenschap aan de culturele, sociale en economische ontwikkeling van de samenleving.



Figuur 1: Organigram van de KNAW

De KNAW ontleent haar gezag aan haar op wetenschappelijke kwaliteit geselecteerde leden. Zij zijn wetenschappers van naam en alle Nederlandse universiteiten zijn hierbij vertegenwoordigd. De circa 500 leden[2] van de KNAW representeren samen een breed spectrum van wetenschappelijke disciplines. De leden kiezen uit hun midden het bestuur, dat op zijn beurt de president benoemt. De huidige president is prof. dr. Robbert Dijkgraaf.

2.2 De afdeling Onderzoek Informatie

De afdeling Onderzoek Informatie is het afgelopen jaar organisatorisch verhuisd van het Bureau naar het instituut DANS. Onderzoek Informatie (8 FTE) houdt zich bezig met het toegankelijk en transparant maken van Nederlands publiek gefinancierd onderzoek en het zichtbaar maken van kennis over onderzoekers en onderzoeksinstituten. Een van de manieren waarop dat gebeurt, is door middel van het product NARCIS.

Elke universiteit en veel onderzoeksinstituten bewaren een groot deel van hun wetenschappelijke output[3] en de bijbehorende metadata in een repository onder de verantwoordelijkheid van de repository managers.

NARCIS leest dagelijks de repositories van de dertien Nederlandse universiteiten en een groot aantal onderzoeksinstellingen uit. Bovendien heeft NARCIS toegang tot de Nederlandse Onderzoek Databank. Hier is onder andere al het onderzoek dat gefinancierd is met NWO-geld, aangemeld. Zo kunnen de gegevens uit die gegevensverzamelingen gecombineerd worden.

NARCIS is ontwikkeld door SURFfoundation en na oplevering door de Beleidsgroep Innovatie Kennisinfrastructuur (BIK) overgedragen aan de KNAW, aangezien de KNAW

2 http://www.knaw.nl/Pages/DEF/26/157.bGFuZz1OTA.html voor een overzicht van de leden
3 publicaties, datasets, patenten etc

overkoepelend opereert. De BIK bestaat uit beleidsmakers van de Nederlandse universiteiten, de KNAW, NWO en de Koninklijke Bibliotheek. Dit samenwerkingsverband waarbij alle universiteiten dezelfde metadata in hetzelfde formaat leveren, is uniek binnen Europa.

2.3 De NARCIS Suite

NARCIS bestaat uit de volgende producten die in de volgende subparagrafen worden toegelicht:

1. Nederlandse Onderzoek Databank (NOD)
2. NARCIS Index
3. NARCIS Portal
4. NARCIS Repository

2.3.1 Nederlandse Onderzoek Databank

De Nederlandse Onderzoek Databank (NOD) is de plek waar al het onderzoek gefinancierd met overheidsgeld, wordt geregistreerd. De gegevens die worden opgeslagen bevatten ondermeer het onderwerp, de duur, de namen van de onderzoekers. Bovendien staan in de NOD alle organisaties (universiteiten, instituten) die zich met wetenschappelijk onderzoek bezighouden en de betrokken onderzoekers.

Eens per jaar geeft de Sdu het rapport *Universiteiten en Onderzoeksinstellingen in Nederland* uit, een neerslag van de data in papieren formaat. In de wandelgangen 'het rode boek' genaamd.

2.3.2 NARCIS Index/Storage

De NARCIS Index is het aggregaat van de repositories van de universiteiten, een aantal onderzoeksinstellingen en de NOD.

De NARCIS Index handelt de zoekvragen van de NARCIS Portal af. Deze uitwisseling gaat via het *Search/Retrieve via URL (SRU)* protocol. Sinds kort is de Index hiermee ook benaderbaar door derden.

2.3.3 NARCIS Portal

De Portal is de showcase voor de Index. De website wordt gebruikt door 'het publiek' dat op zoek is naar Nederlands onderzoek. Belangrijker is de functie die de website heeft met betrekking tot de repository managers en de beleidsmakers. De Portal toont wat er verscholen zit in de data. De Portal geeft de relaties tussen onderzoek, onderzoekers en publicaties weer. Binnen deze afstudeeropdracht breid ik dit verder uit. Naarmate de repositories meer typen data aanbieden, zoals de Digital Author Identifier (DAI)[4], kunnen de dwarsverbanden beter gelegd worden.

4 Zie 10.1 Afstudeerplan

2.3.4 NARCIS Repository

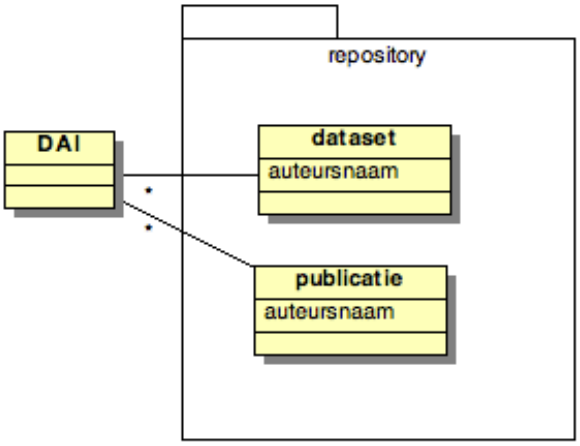
Internet is een vluchtig medium geworden. Om duurzaam aan digitale objecten (zoals publicaties) te refereren, volstaat de URL niet langer. Om na verhuizing of ophef van domeinen niet met de gevreesde http error *code 404* geconfronteerd te worden, is een andere referentie nodig. Hiervoor is de Digital Object Identifier (DOI) in het leven geroepen. In Nederland geïmplementeerd door de Persistent Identifier. Deze wordt in de metadata van een digital object opgeslagen. Door middel van een resolver kan deze DOI vertaald worden naar een URL, analoog aan een telefoonboek. Men verwijst dus naar de DOI en de resolver in plaats van direct de URL te noemen.

De NARCIS Repository kan als input voor de resolver fungeren door de DOI te koppelen aan een URL. Daarnaast kan de Repository ook internationaal gebruikt worden als Nationale Repository, bijvoorbeeld door het DRIVER[5] project.

5 Digital Repository Infrastructure Vision for European Research

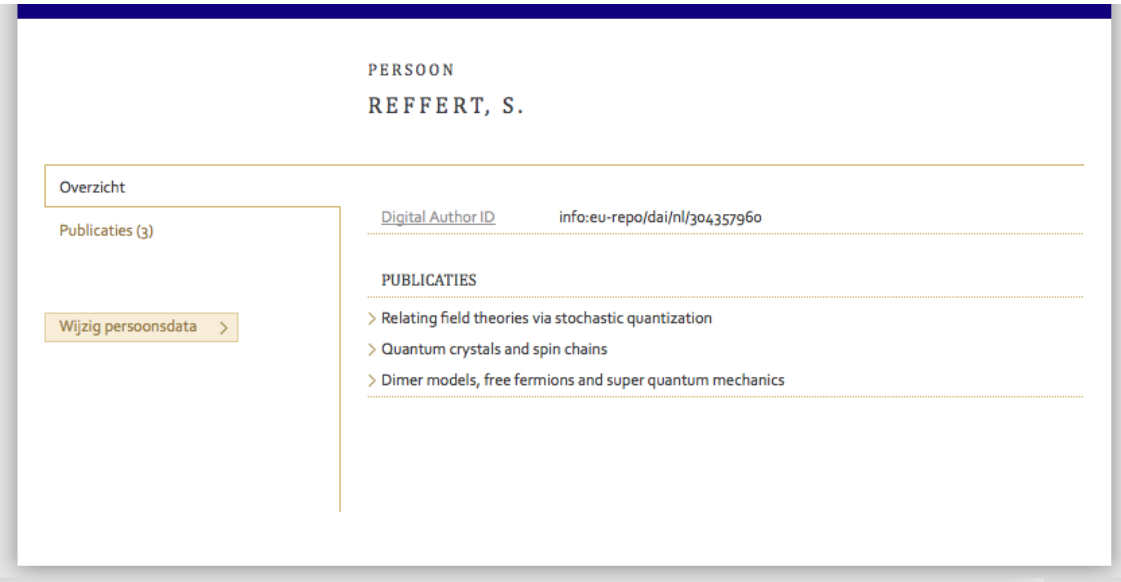
3 Probleemstelling

Van oudsher is de auteursnaam een attribuut van de metadata van een publicatie. De auteursnaam was dus geen zelfstandig object maar vrije tekst. Daardoor zijn er in verschillende publicaties verschillende schrijfwijzen van auteursnamen en incomplete initialen in de repositories terug te vinden. Om een auteur eenduidig te kunnen identificeren is een aantal jaar geleden de DAI ingevoerd, de Digital Author Identifier. Langzaam maar zeker vindt die zijn weg in de repositories. Hierdoor is het nu wel mogelijk alle publicaties van een specifieke auteur te vinden.



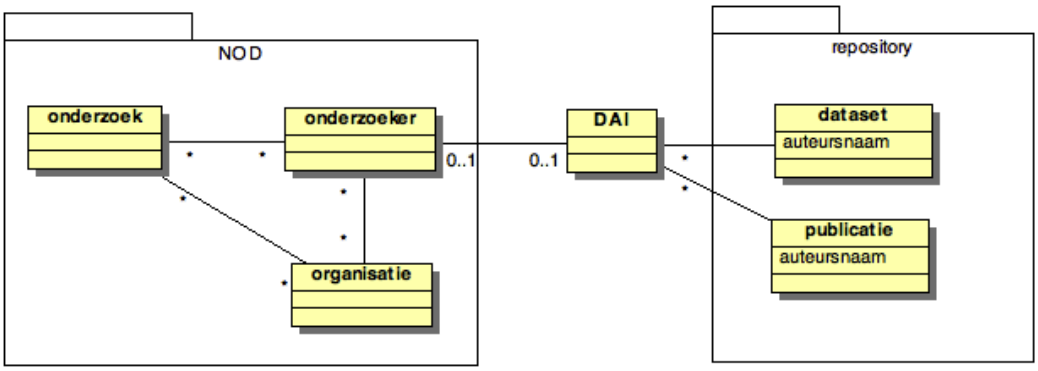
Figuur 2: de DAI als verbindende factor tussen repositories

Om te laten zien wat een koppeling van de data uit de verschillende repositories middels de DAI op kan leveren, is een portal gebouwd: www.narcis.nl. Vanuit dit centrale punt kan de data uit de verschillende repositories met één zoekopdracht worden doorzocht.
De dwarsverbanden tussen de objecten worden door de portal nog niet inzichtelijk gemaakt. Op dit moment is het wel mogelijk op DAI-nummer een persoonspagina te maken, waarop de publicaties van die auteur getoond worden.



Figuur 3: Persoonspagina van de auteur met DAI-nummer 304357960
http://www.narcis.nl/person/info:eu-repo/dai/nl/304357960

De afdeling Onderzoek Informatie wil graag dat instituten hun repositories op meer punten voorzien van DAI data. Voor medewerkers van een universiteit of instituut is het over het algemeen gemakkelijk om van eigen onderzoekers de DAI te achterhalen en in de metadata op te nemen. De DAI van onderzoekers van andere instituten die betrokken zijn bij een publicatie, is lastig om te achterhalen en toe te voegen. Daarom wordt deze data niet vermeld. De repository managers kunnen de DAI wel opzoeken bij OCLC-PICA[6].



Figuur 4: de DAI als verbindende factor tussen de repositories en de NOD

Binnen de NOD kan bij een onderzoeker zijn persoonlijke DAI opgeslagen worden. Hierdoor wordt een auteur geïdentificeerd als een bij de NOD geregistreerde onderzoeker. Zo vormt de DAI niet alleen de verbindende factor tussen de repositories onderling, maar ook tussen de repositories en de NOD.

6 Zie bijlage 10.2 Afkortingenlijst

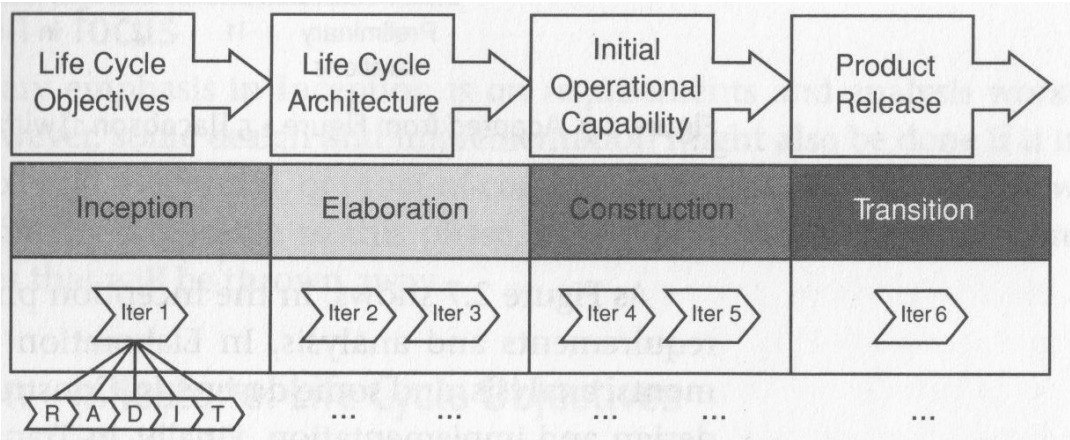
Figuur 5: Persoonspagina van professor dr. H.J. Bennis. Naast gegevens uit de NOD zijn ook 144 publicaties uit de repositories opgenomen.

Figuur 5 toont de pagina voor een specifieke onderzoeker. De getoonde data komt deels uit de NOD en deels uit de repositories.

De opdrachtgever van de afstudeeropdracht wil door deze dwarsverbanden te visualiseren de meerwaarde van de DAI aantonen. De opdracht valt in twee delen uiteen. In het eerste deel onderzoek ik de mogelijke dwarsverbanden tussen de NOD en de repositories. Het tweede deel bestaat uit het ontwerpen en implementeren van een tweetal demonstrators om deze dwarsverbanden te illustreren.

4 Te gebruiken methoden en technieken

Aangezien een groot deel van de opdracht bestaat uit het onderzoek, wil ik een software ontwikkelmethode gebruiken waarbij dat onderzoek als aparte fase zichtbaar is. Ik gebruik hiervoor de inception fase van het Unified Process. Het doel van inception fase is immers het overtuigen dat een systeem meerwaarde biedt. Normaal gesproken wordt bij de business case het profijt uitgedrukt in geld. In dit onderzoek wordt dit uitgedrukt in draagvlak bij de stakeholders. Bovendien worden de generieke functionele requirements en risico's aangekaart. Het resultaat van deze inception fase is een onderzoeksrapport [TJR01] met een lijst van potentiële toepassingen. Deze worden gerangschikt volgens een bepaalde beslissingstechniek.



Figuur 6: de UP fases met de vijf disciplines Requirements, Analysis, Design, Implementation en Test. Uit: [M&T01]

Tijdens het tweede deel van de afstudeeropdracht worden twee demonstrators ontwikkeld. Voor beide is in de elaboration en construction fase een aparte iteratie.

De UP-fasering van het totale project is als volgt:

Deel	Fase	artefacten
onderzoek	Inception	• onderzoeksrapport [TJR01] • glossary [TJR05] • generieke functionele requirements van de demonstrators
ontwerp	Elaboration	Iteratie demonstrator I • functionele requirements specifiek voor demonstrator I [TJR03] • prototypes • analysis model • design model
		Iteratie demonstrator II • functionele requirements specifiek voor demonstrator II [TJR06] • analysis model • design model • prototype
	Construction	Iteratie demonstrator I • testcode • code
		Iteratie demonstrator II • testcode • code
	Transition	• evaluatie • aanbevelingen

Tabel 1: fasering van de afstudeeropdracht

Tijdens het ontwikkelen is gewerkt volgens Test Driven Development (TDD) [M&T03], een methode uit de Agile stal. TDD gaat uit van het principe dat het beschrijven van de functionaliteit van een klein deel van het systeem gemakkelijker is dan van een groot deel. Het legt deze functionaliteit vast in een test, die in eerste instantie faalt. Hierna wordt de benodigde code geproduceerd om de test te laten slagen en de code wordt gerefactored [M&T06] om functionele overlap te verwijderen.

Deze continu uitgevoerde tests weerspiegelen de progressie en geven het vertrouwen dat met het refactoren en wijzigen van de code niet ongemerkt bugs worden geïntroduceerd. Bovendien nodigt het uit tot het schrijven van een systeem met 'high cohesion en loose coupling'. Twee begrippen die bij een object georiënteerde taal als Java van belang zijn.

Nadat de demonstrators ontwikkeld zijn, wordt in de transition fase onderzocht en beschreven welke stappen nog genomen moeten worden om het systeem klaar te maken voor de productieomgeving. Dit betreft niet slecht technische stappen. Ook bedrijfsmatige en / of financiële beslissingen spelen hier een rol.

5 Onderzoeksfase

In gesprekken met de opdrachtgever en bedrijfsmentor is de onderzoeksvraag als volgt geformuleerd.

> *Hoe kunnen de mogelijkheden en de meerwaarde van de Digital Author Identification gedemonstreerd worden op een manier die de verschillende stakeholders aanspreekt?*

5.1 De stakeholders

In de onderzoeksvraag wordt uitdrukkelijk gesproken over de stakeholders. Ik heb drie groepen stakeholders geïdentificeerd:

1. projectleiding van NARCIS
   de projectleiding heeft de ambitie van NARCIS het portal te maken voor alle vragen met betrekking tot lopend onderzoek en wetenschappelijke output. Hiervoor is het correct en volledig toevoegen van de DAI cruciaal. Bovendien zouden de demonstrators opgenomen kunnen worden in de webportal.
2. beleidsmakers van de universiteiten
   De beleidsmakers van de universiteiten zijn vertegenwoordigd in de BIK en bepalen de koers die de universiteiten varen met betrekking tot het publiceren van de onderzoeksgegevens. In mei 2005 hebben de eerste Nederlandse universiteiten de Berlin Declaration on Open Access ondertekend. Om open access te faciliteren wordt gebruik gemaakt van repositories om de producten en de metadata op te slaan. Deze metadata wordt via NARCIS beschikbaar gesteld. Het invoeren van de DAI is hier een onderdeel van.
3. repository managers
   De repository managers zijn verantwoordelijk voor de repository van hun instituut. De repositories zijn in eerste instantie bedoeld voor de opslag van de wetenschappelijke output. De bijbehorende metadata wordt aangeboden aan het indexeringproces van NARCIS[7]. De repository managers kunnen aan de medewerkers van de bibliotheek tools aanbieden die het invoeren van DAI-nummers bij auteurs vergemakkelijkt.

Ik heb gesproken met vertegenwoordigers van de drie groepen om zicht te krijgen op de criteria van de verschillende stakeholders.

Ad 1. Als eerste heb ik een oriënterend gesprek gehad met dhr. Chris Baars, BA, de projectleider van NARCIS en mijn bedrijfsmentor tijdens dit afstudeerproject. We hebben een voorlopige lijst met criteria opgesteld, die ik als input voor de overige gesprekken heb gebruikt.

Ad 2. Als vertegenwoordiger van de beleidsmakers heb ik gesproken met drs. Elly Dijk. Zij was lid van de BIK, de Beleidsgroep voor Innovatie en Kennisinfrastructuur. Hierin zijn

7 Zie bijlage 10.12 Harvesten

informatiemanagers en beleidsmakers van de universiteiten vertegenwoordigd.

Ad 3. Als vertegenwoordiger van de repository managers heb ik gesproken met drs. Arjan Hogenaar en Armand Guicherit, repository managers van de KNAW en lid van de Werkgroep Repository Managers (WRM). Dhr. Hogenaar is daarnaast betrokken bij een project over persistent identifiers, een broertje van de DAI voor publicaties (Digital Object Identifier)[8]. Bovendien heb ik vergaderingen van de WISH[9] bijgewoond en notulen doorgenomen.

De verslagen van deze gesprekken heb ik naar de betrokkenen gestuurd ter aanvulling en goedkeuring.

5.2 De criteria

Het bleek lastig om over de criteria te spreken. De gesprekspartners wilden graag vertellen welke ideeën ze over NARCIS hebben. Zij spraken liever over oplossingen dan over de achterliggende problemen en wensen. Ik heb hun oplossingen opgesplitst in de achterliggende criteria. Ik heb aan de hand van hun ideeën met hen gesproken over de aspecten daarin die ze interessant vonden. Op die manier kwamen de criteria alsnog boven water.

Uit deze gesprekken[10] heb ik de criteria gedestilleerd. Deze heb ik gegroepeerd en gegeneraliseerd om ervoor te zorgen dat de criteria:

1. orthogonaal,
2. terzake,
3. compleet
4. en kwantificeerbaar zijn.

Ad 1. Om de alternatieven eerlijk te kunnen vergelijken, moeten de criteria orthogonaal zijn. Ze mogen niet dubbel voorkomen, elkaar overlappen of tegengesteld zijn.

Ad 2. De criteria moeten daadwerkelijk met het probleem te maken hebben. Pas dan kan vastgesteld worden in hoeverre een alternatief een oplossing biedt.

Ad 3. Om vast te stellen of de meest belangrijke criteria benoemd zijn, heb ik alle oplossingen die de stakeholders aandroegen, teruggebracht tot hun achterliggende criteria. Toen ik geen nieuwe criteria tegenkwam, en de stakeholders ook geen andere oplossingen noemden, heb ik geconstateerd dat de set compleet is.

Ad 4. Van elk alternatief moet aangegeven kunnen worden in hoeverre het aan een criterium voldoet. Deze moeten dus kwantificeerbaar zijn om de alternatieven onderling te kunnen vergelijken. Bij het ene criterium is dit eenvoudiger dan bij het andere. Soms zal een educated guess moeten voldoen.

8 Zie 2.3.4 NARCIS Repository
9 Zie 10.2 Afkortingenlijst
10 Zie 10.4 Gespreksverslagen stakeholders

De criteria zijn hier kort omschreven en voorzien van een werktitel, zie [TJR01] voor een bredere behandeling.

C1. hoog DAI gehalte
Wordt de demonstrator beter naarmate er meer DAI's in de repositories en NOD beschikbaar zijn?

C2. Onderling verbinden van repositories
Dit is een ja/nee criterium

C3. geen onjuiste data weergeven
Wanneer de data geïnterpreteerd moet worden (bijvoorbeeld om van twee publicaties vast te stellen of het dezelfde betreft), hoe groot is dan de kans op het tonen van onjuiste data?

C4. geen incomplete weergave
hoe groot is de kans dat alle data die beschikbaar is, daadwerkelijk gevonden wordt? Dit gaat niet over de inherente incompleetheid van de repositories. Dit aantonen is juist een onderdeel van dit project.

C5. Hoge 'wow' factor
In welke mate inspireert het mensen? Levert het nieuwe dwarsverbanden op die anders niet (makkelijk) gezien zouden worden?

C6. kleine kans op mislukking
Hoe groot is de kans dat de toepassing mislukt, anders dan door het lage aantal DAI's in de repositories?

C7. grote betrokkenheid NOD
In welke mate is de NOD noodzakelijk voor dit alternatief?

C8. auteur staat centraal
Zorgt dit alternatief ervoor dat de informatie van individuele onderzoekers / auteurs beter/mooier/uitgebreider getoond wordt?

C9. Privacy
Niet alle data die tot onze beschikking staat, mag gepubliceerd worden. Bijvoorbeeld de data van OCLC-PICA is intern wel bekend, maar mag niet voor alles gebruikt worden. Dit is een ja/nee criterium

C10. Instituut staat centraal
Zorgt dit alternatief ervoor dat de informatie voor een specifiek instituut beter tot zijn recht komt?

Aangezien mijn gesprekspartners druk bezette agenda's hebben, heb ik een methode gekozen waarbij ik hen individueel kon spreken en de tijd beperkt bleef tot twee gespreksronden. De 'gewogen criteria'-methode is een kardinale[11] methode om alternatieven met elkaar te vergelijken. De stakeholders spreken afzonderlijk hun voorkeuren uit. Gebaseerd hierop krijgt elk criterium voor iedere stakeholder een gewicht.

11 kardinaal (op de grootte betrekking hebbend) in tegenstelling tot ordinaal (op de rangorde betrekking hebbend), zoals Pugh

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	totaal
C1	-	1	0	1	1	1	0	1	0	1	**6**
C2	0	-	0	1	1	1	1	1	0	1	**6**
C3	1	1	-	1	1	1	1	1	1	1	**9**
C4	0	0	0	-	1	0	0	1	1	1	**4**
C5	0	0	0	0	-	1	0	1	0	1	**3**
C6	0	0	0	1	0	-	1	1	0	1	**4**
C7	1	0	0	1	1	0	-	1	0	1	**5**
C8	0	0	0	0	0	0	0	-	0	1	**1**
C9	1	1	0	0	1	1	1	1	-	1	**7**
C10	0	0	0	0	0	0	0	0	0	-	**0**

Tabel 2: uitkomst van de 'gewogen criteria'-methode: de criteria gewogen volgens de NARCIS projectleiding uit [TJR02]

In tabel 2 is als voorbeeld opgenomen de uitkomst van dit proces voor de stakeholder *NARCIS projectleiding*. Wanneer het horizontale criterium belangrijker gevonden wordt dan het verticale, dan krijgt het horizontale criterium 1 punt. Zo betekent de eerste 1 op de eerste regel dat de NARCIS projectleiding criterium C1 *hoog DAI gehalte* belangrijker vindt dan C2 *onderling verbinden van repositories.* De onderdriehoeksmatrix is de inverse van de bovendriehoeksmatrix. De laatste kolom toont de som voor dit criterium, het uitgangspunt voor de weegfactor.

Op deze manier zijn de weegfactoren voor de drie stakeholders samengesteld.

5.3 De alternatieven

Uitgaande van deze criteria heb ik een brainstormsessie georganiseerd met het NARCIS ontwikkelteam en de repository manager van de KNAW om ideeën voor demonstrators te bedenken. Ook de ideeën uit de eerste gespreksronde zijn meegenomen.

A1.	Woordenwolk samenstellen bij persoon
A2.	Extra bronnen raadplegen
A3.	Experts vinden
A4.	Profielpagina's koppelen
A5.	Coworkers tonen
A6.	Academische loopbaan tonen
A7.	Onderzoek aan publicatie koppelen
A8.	Publicaties bij instituten tonen
A9.	Emerging topics identificeren
A10.	Overlappen tussen onderzoeksgebieden vaststellen
A11.	Verschuivende interesse-gebieden tonen

Tabel 3: Shortlist met werktitels van de alternatieven. Zie [TJR01] voor de volledige beschrijving.

De veelheid aan ideeën heb ik teruggebracht tot een shortlist van alternatieven, rekening houdend met de veronderstelde technische beperkingen en de tijdslimiet van het afstudeerproject. Alle ideeën waarbij de rol van de DAI te marginaal was heb ik niet meegenomen.

5.4 De uitkomst

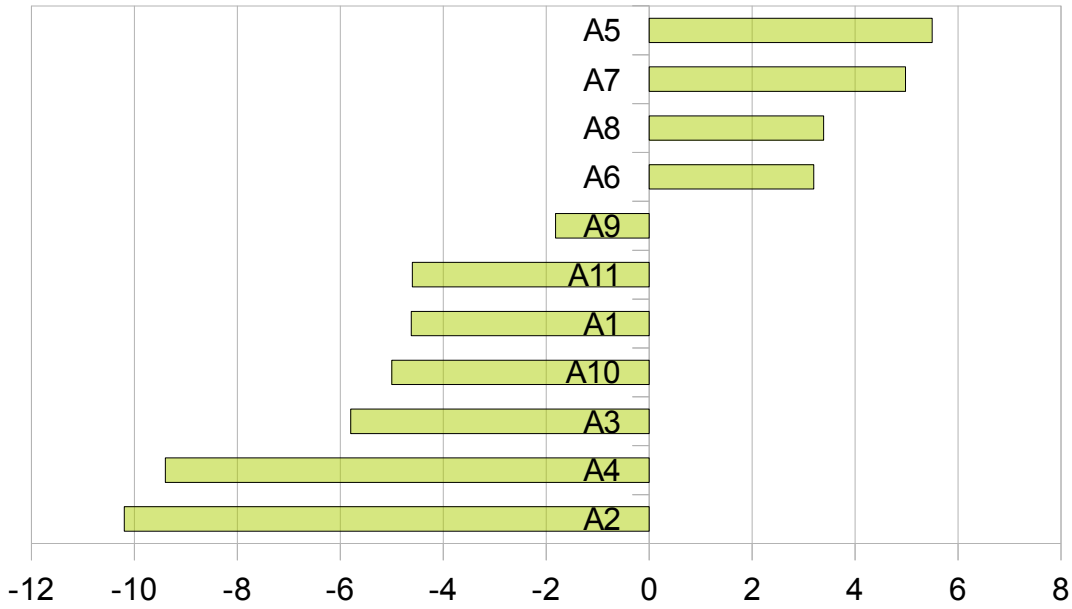
Aangezien de criteria orthogonaal en compleet zijn, spannen ze een n-dimensionale ruimte op. Door aan de alternatieven voor elk criterium een mate van overeenkomst toe te kennen, kunnen ze gerepresenteerd worden als vectoren in deze n-dimensionale ruimte. Alle stakeholders hebben voor elk tweetal criteria hun voorkeur aangegeven, zie tabel 2. Hiermee claimen ze een deel van de oplossingsruimte als favoriet. Alternatieven die zich in dit deel bevinden hebben de voorkeur van deze stakeholder. Om de alternatieven voor de stakeholders te kunnen rangschikken, heb ik ze gescoord op de criteria, zie tabel 4, genomen uit [TJR01].

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
A1	50	1	60	100	66	70	100	1	0	0
A2	10	0	70	0	0	40	50	1	1	0
A3	10	1	40	100	100	40	100	0	0	0
A4	10	0	70	0	0	20	50	1	1	0
A5	100	1	0	50	100	0	0	1	0	0
A6	100	1	10	0	100	0	50	1	1	50
A7	100	1	0	0	66	0	100	0	0	0
A8	50	0	50	0	33	40	100	0	0	100
A9	100	1	0	100	66	70	100	0	0	0
A10	100	1	30	100	100	40	100	0	0	50
A11	100	1	60	100	100	70	50	1	0	0

Tabel 4: de alternatieven gescoord volgens de criteria

De som van alle producten van zulke {criterium, alternatief} paren levert voor elke stakeholder per alternatief een getal. Deze kunnen vergeleken worden om het alternatief met de meeste draagvlak te identificeren.

De weegfactor die de stakeholders aan een criterium toekennen, wordt vermenigvuldigd met de score van de alternatieven op dat criterium. Dit levert voor elke stakeholder een lijst met alternatieven, gesorteerd op voorkeur. Figuur 7 toont de uitkomst voor de beleidsmakers van de universiteiten.



Figuur 7: de volgorde van de alternatieven volgens de weegfactoren van de beleidsmakers

Op basis van de uitkomsten voor de verschillende stakeholders heb ik, met fiat van de opdrachtgever, A5 (c*oworkers tonen*) en A7 (o*nderzoek aan publicatie koppelen*) gekozen om daadwerkelijk als demonstrator te ontwikkelen.

6 Demonstrators

De onderzoeksfase is afgesloten met een keuze voor de volgende twee demonstrators.

Demonstrator I : coworkers tonen

"*Tonen van de (nationale) academische omgeving van een persoon: de co-auteurs en co-onderzoekers. Dit kan bijvoorbeeld chronologisch, in een graph, alfabetisch. Bovendien zou getoond kunnen worden op welk gebied de samenwerking plaatsvond door middel van termen uit de titel en samenvatting van de publicatie/onderzoek en de NOD-classificatie.*" [TJR01]

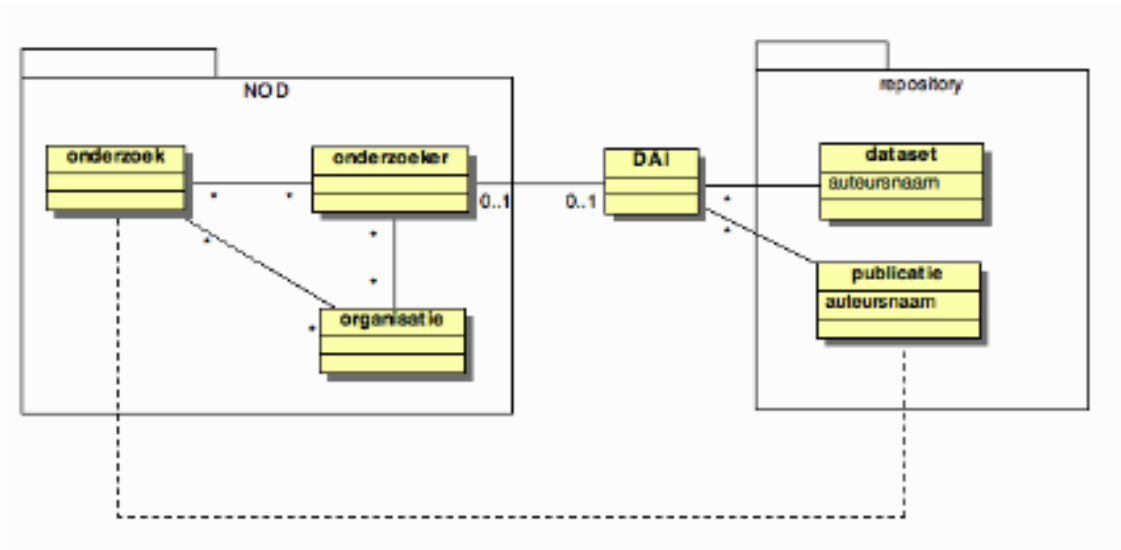
Citaat 1: Demonstrator I: coworkers tonen

Demonstrator II: onderzoek aan publicatie koppelen

"Het idee is om onderzoeken uit de NOD te koppelen aan publicaties en datasets. Er kan gematched worden op titel, samenvatting, auteur, etc. Bijvoorbeeld: onderzoekers die gekoppeld zijn aan een onderzoek in de NOD, en samen een artikel hebben gepubliceerd ten tijde van dat onderzoek." [TJR01]

Citaat 2: Demonstrator II: onderzoek aan publicatie koppelen

Een significant verschil tussen de beide demonstrators is dat demonstrator I *coworkers tonen* bestaande verbanden visualiseert, waar demonstrator II *onderzoek aan publicatie koppelen* een stap verder gaat en nieuwe relaties legt. Deze niet direct aanwezige koppeling komt tot stand via de DAI. De kans dat een publicatie aan het verkeerde onderzoek gekoppeld wordt, moet geminimaliseerd worden.



Figuur 8: demonstrator II: het leggen van een nieuwe relatie tussen onderzoek en publicatie, vergelijk figuur 4

6.1 Elaboration fase

In de elaboration fase is het probleemdomein van de beide demonstrators geanalyseerd en een design model opgesteld.

6.1.1 Iteratie demonstrator I: coworkers

Analyse van het probleem

Om de samenwerkende auteurs rondom een persoon bij een persoonspagina te kunnen bepalen, moeten dit dus unieke, identificeerbare objecten zijn. Maar de auteursnaam is 'slechts' als een attribuut vermeld binnen de metadata van een publicatie en niet als een afzonderlijke entiteit. Tot voor kort werd de auteur aangeduid met vrije tekst als `<dc:creator>vrije tekst</dc:creator>` in de metadata van een publicatie. Repositories die zich aan de laatste afspraken confirmeren[12], hebben een uitgebreidere set predicaten tot hun beschikking.

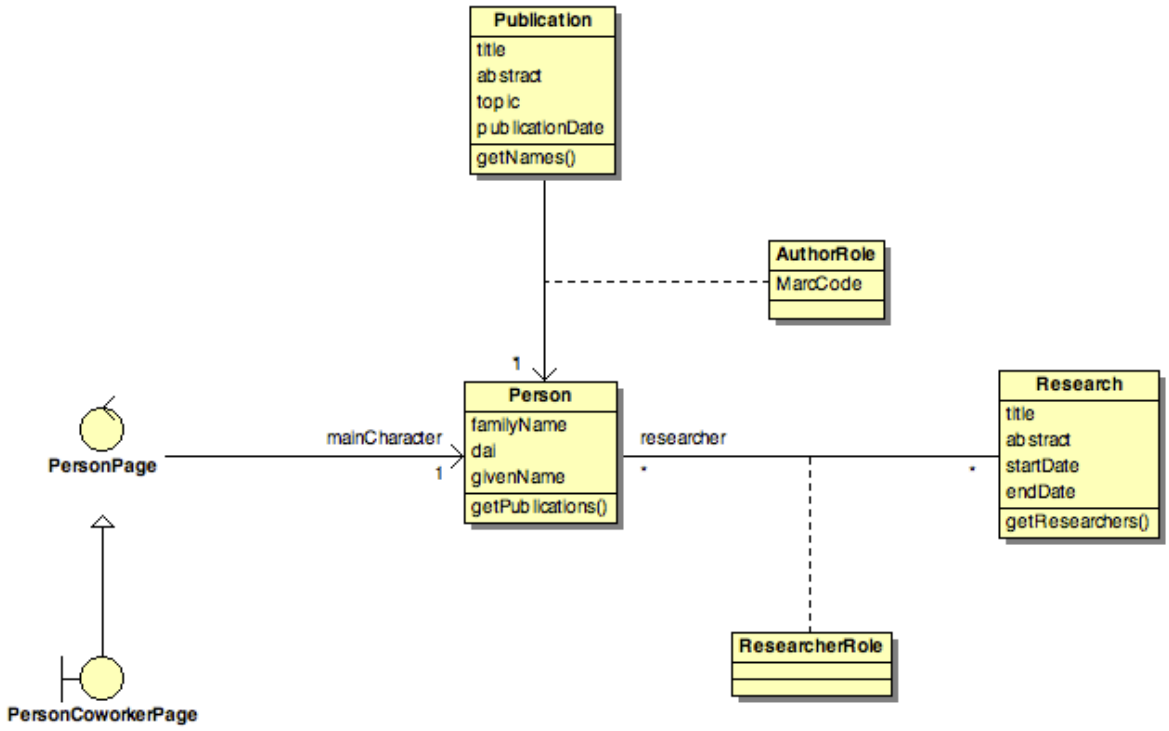
```
<mods [...]>
  <titleInfo><title>Onmisbaar water</title></titleInfo>
  <dc:creator>L.J. de Haan</dc:creator>
  [...]
</mods>

<mods [...]>
  <titleInfo><title>Onmisbaar water</title></titleInfo>
  <name type="personal" ID="id2856079">
    <namePart type="family">Haan, de</namePart>
    <namePart type="given">L.J.</namePart>
    <role>
      <roleTerm type="code" authority="marcrelator">aut</roleTerm>
    </role>
  </name>
  <extension>
    <daiList schemaLocation="info:eu-repo/dai [...]">
      <identifier IDref="id2856079" authority="info:eu-repo/dai/nl">072791497</identifier>
    </daiList>
  </extension>
  [...]
</mods>
```

Citaat 3: de naam van de auteur binnen een mods-component. Bovenste fragment toont de 'oude' manier, onder staat de 'nieuwe' manier.

Citaat 3 toont een voorbeeld van beide notaties. Naast een aparte voor- en achternaam, kan ook de rol van de persoon (of instelling) aangegeven worden in MARC-code[13] en natuurlijk de DAI, gerealiseerd in een mods-extensie. Deze geeft de auteur zijn identificatie, waardoor eenzelfde persoon ook elders herkend kan worden en de auteur een entiteit wordt.

12 http://wiki.surffoundation.nl/display/standards/Use+of+MODS
13 http://www.loc.gov/loc.terms/relators/



Figuur 9: class diagram uit het analysis model

Figuur 9 toont een class diagram uit het analysis model. Hierin staan de classen zoals die hierboven zijn beschreven. De PersonCoWorkerPage geeft de plek aan binnen de NARCIS Portal waar deze data getoond zou moeten worden. De Person die een author is van een Publication, is, logisch gezien, van dezelfde class als een researcher bij een Research. Maar ze zijn afkomstig uit andere bronnen. De match moet gemaakt worden met behulp van de DAI.

Afgezien van de herkomst van de data moet in deze fase ook vastgesteld worden hoe deze gevisualiseerd gaat worden[14]. De requirements vloeiden grotendeels voort uit de brainstormsessie die ik gehouden heb met het NARCIS ontwikkelteam. Enkele belangrijke requirements hadden betrekking op de overzichtelijkheid van de visualisatie. Ik heb een aantal voorbeeldcases geconstrueerd, waaronder een casus die veel zal voorkomen en enkele die voor veel visualisaties een worst-case vormen. Door middel van prototypes zijn van deze cases verschillende visualisaties gemaakt. Hierdoor konden de sterke en zwakke punten van de verschillende typen visualisaties vergeleken worden, wat het keuzeproces structureerde. Het resultaat van dit keuzeproces heb ik gepresenteerd aan het NARCIS ontwikkelteam en projectleiding, zodat allen op de hoogte waren van de geconstateerde voordelen en nadelen [TJR08].

Een van de voorbeeldcases bestaat uit de volgende publicaties van Aalbers, met elk een set auteurs:

- pub1: Bladvlekkenziekte weer toenemend probleem door vochtiger klimaat (2010) S1: {Aalbers, Deurloo, Mustert, Kloosterman}
- pub2: Pythium kan voor ernstige problemen zorgen (2007)

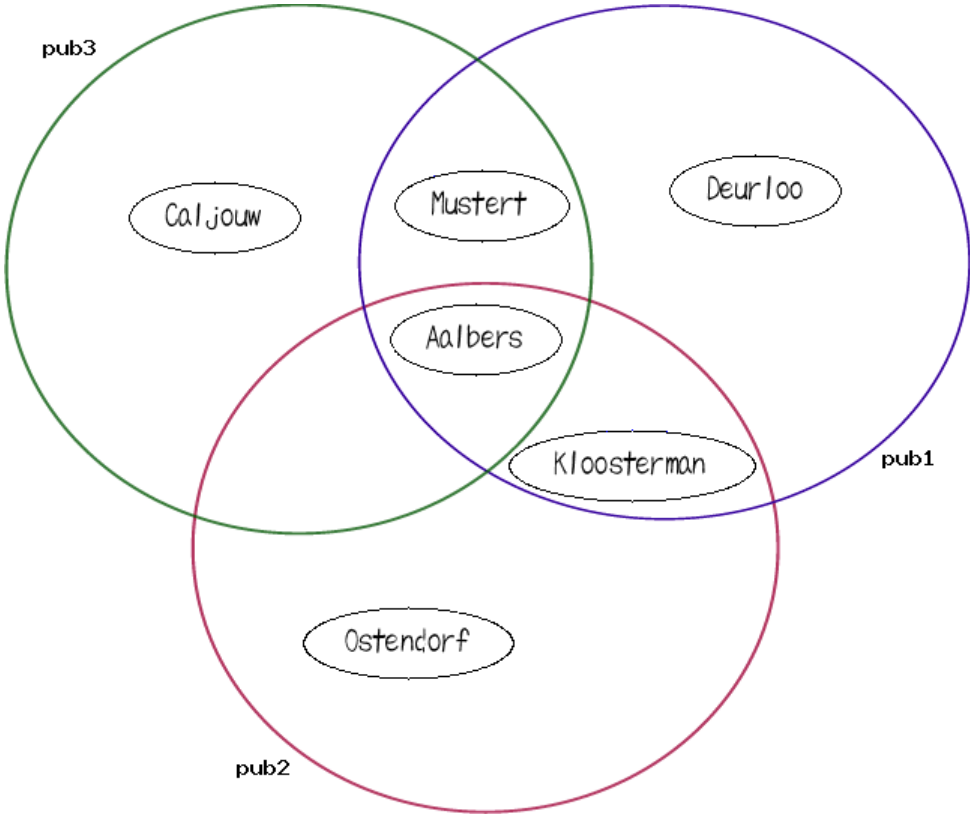
14 Zie 10.7 Uitwerking demonstrator Coworker

S2: {Aalbers, Kloosterman, Ostendorf}
- pub3: Opsporen oorzaak zwarte spikkels in radijs (2006)
  S3: {Aalbers, Mustert, Caljouw}

Op de PersonCoWorkerPage van Aalbers zou dit als volgt weergegeven kunnen worden

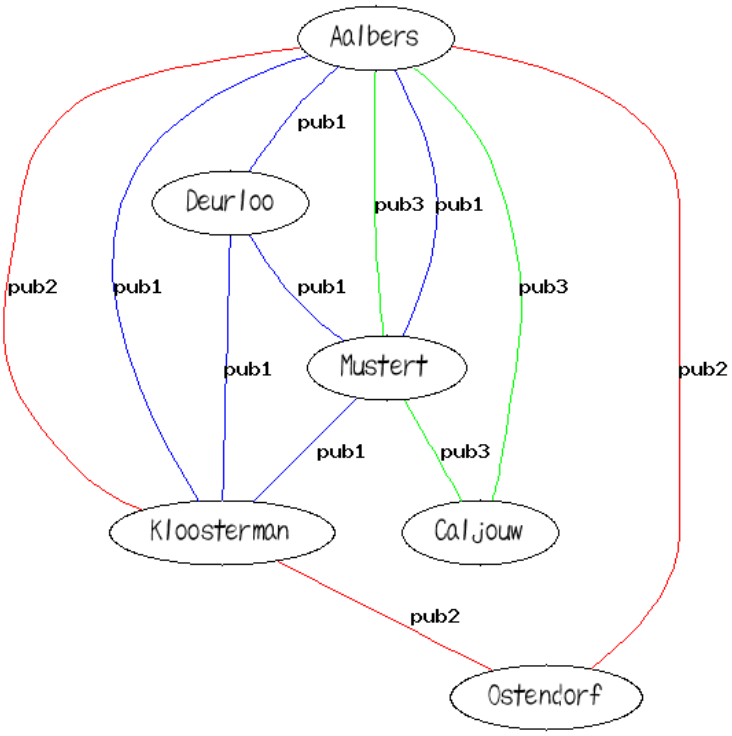
Caljouw - Deurloo - Kloosterman (2) - Mustert (2) - Ostendorf

De meest voor de hand liggende visualisatie van publicaties is een Venn diagram. Elke publicatie wordt weergegeven als een set van auteurs.



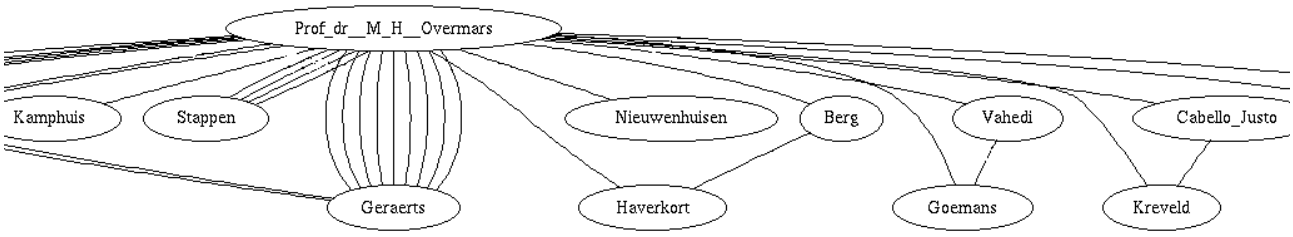
Figuur 10: Venn diagram met de drie publicaties pub1 (blauw) , pub2 (rood) en pub3 (groen)

Venn diagrammen geven overlappen van sets goed weer, maar zijn beperkt in de hoeveelheid sets. Al vanaf vijf sets wordt het lastig te zien welke elementen tot welke set behoren. Venn diagrammen zijn voor dit probleem dus niet bruikbaar, twintig publicaties voor een auteur is niet ongewoon.
Wanneer de sets niet als gebieden getoond worden, maar als verbindende edges, kan zo'n netwerk als graph weergegeven worden. Figuur 11 toont zo'n graph, gemaakt met GraphViz. Hierin zijn de afzonderlijke publicaties te herkennen als gekleurde edges. De nodes zijn de auteurs.



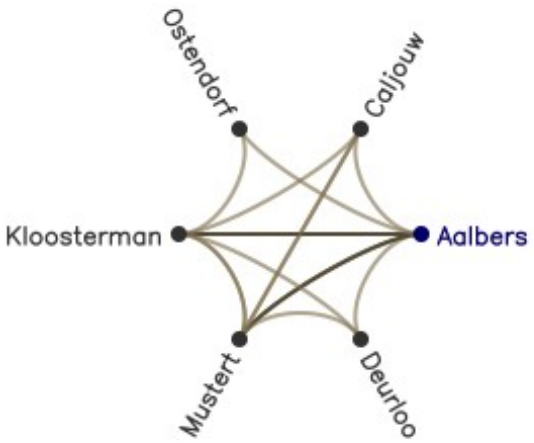
Figuur 11: De relaties pub1 (blauw) , pub2 (rood) en pub3 (groen)

Het grote nadeel van deze methode is weer het gebrek aan schaalbaarheid. Het is, anders dan bij een Venn diagram, wel mogelijk om een casus met twintig publicaties te tekenen, maar is niet overzichtelijk. Bovendien ziet elke graph er anders uit dan de andere, er is geen *uniformiteit*, een van de requirements. Zoals in figuur 12 getoond wordt, is dit dus een instabiele weergave.



Figuur 12: deel van de omgeving van een auteur met meer publicaties.
Dit wordt snel onoverzichtelijk

De uiteindelijk gekozen visualisatie is een graph, waarbij de nodes in een cirkel geplaatst worden. Dit zorgt voor uniformiteit en maakt het schaalbaar. Figuur 13 toont de voorbeeldcasus. Dubbele edges tussen nodes worden hier vervangen door een anders gekleurde edge. Deze representatie is een zogenaamd moowheel dat gebruik maakt van de javascript library mootools.js. Figuur 14 toont de achterliggende datastructuur.



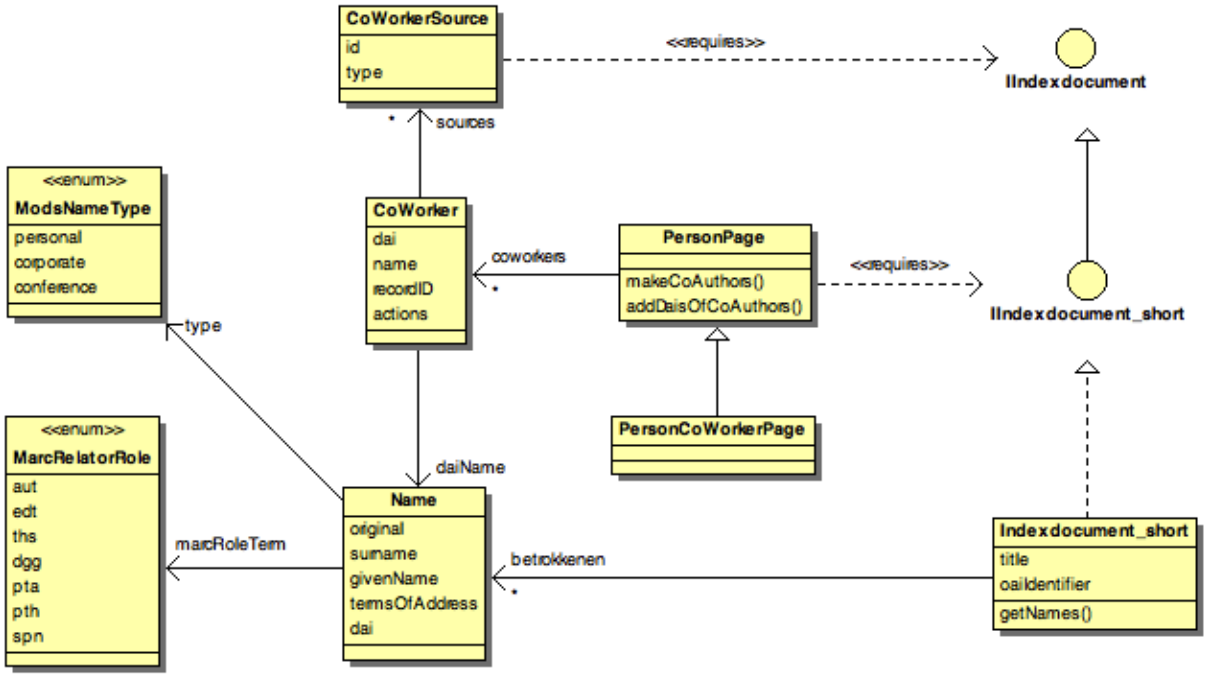
Figuur 13: Visualisatie met de mootools library waarbij de helderheid een indicatie is voor het aantal gezamenlijke publicaties

```
var data =  [
{id: 'Aalbers',
connections: [ ['Aalbers', 1] ]},
{id: 'Deurloo',
connections: [ ['Aalbers', 1] ]},
{id: 'Mustert',
connections: [ ['Aalbers', 2], ['Deurloo', 1], ['Kloosterman', 1], ['Caljouw', 1] ]},
{id: 'Kloosterman',
connections: [ ['Aalbers', 2], ['Deurloo', 1], ['Mustert', 1], ['Osterndorf', 1] ]},
{id: 'Ostendorf',
connections: [ ['Aalbers', 1], ['Kloosterman', 1] ]},
{id: 'Caljouw',
connections: [ ['Aalbers', 1], ['Mustert', 1] ]}
];
```

Figuur 14: versimpelde datastructuur voor het moowheel zoals getoond in figuur 13. Een connection bestaat uit [target, weight]

De volgorde waarin de nodes genoemd worden, is de volgorde waarin ze getekend worden langs de rand van de cirkel. Dit heeft natuurlijk invloed op de uiteindelijke visualisatie. De lengte van de edges heeft geen betekenis.

Aan de hand van het gekozen visualisatie-type is verder gewerkt richting het design-model. Het opgestelde analysis model was voldoende om de structuur (zie figuur 14) te ondersteunen. Figuur 15 toont een bijbehorend class diagram uit het design model.



Figuur 15: class diagram uit het design model. Relatie tussen Person en Publication

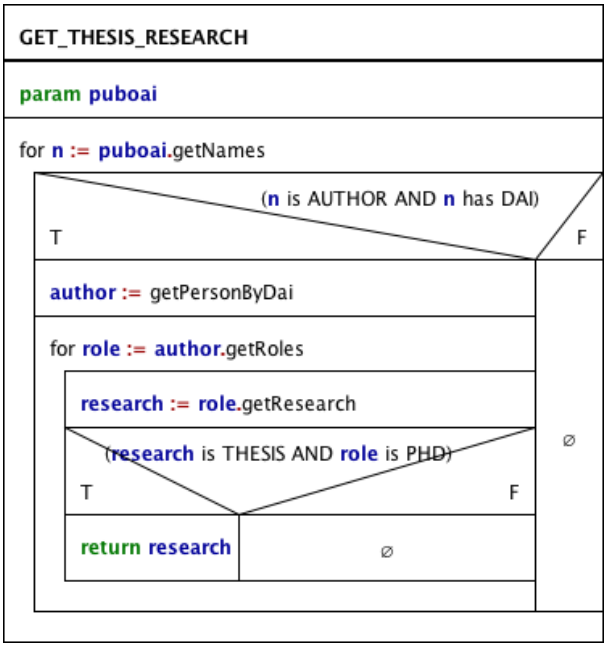
Om aan requirement 4 dat *de visualisatie ingepast kan worden in de bestaande NARCIS webportal* te voldoen, heb ik ervoor gekozen zoveel mogelijk gebruik te maken van bestaande NARCIS libraries. Dit beïnvloedt de richting van het design model. Een aantal benodigde classen is natuurlijk reeds aanwezig. De interface IIndexdocument is het aanspreekpunt voor een Publication. De verschillende formaten waarin de publicaties opgeslagen zijn,[15] hebben elk hun eigen subinterface. Het *short* formaat bevat voor ons probleem voldoende data, namelijk de titel, een identificatie (de oai-identifier) en de betrokkenen in de vorm van Name-objecten.
Het probleem met deze Name-objecten is dat ze elk maar betrekking hebben op één Indexdocument_short. Om auteurs met meer dan één publicatie (lees Indexdocument) te construeren heb ik een nieuwe class geïntroduceerd: de CoWorker. Een CoWorker representeert een auteur, van één of meer publicaties of onderzoeken, zie figuur 15. Dit wordt bijgehouden in de CoWorkerSource.

15 Zie bijlage 10.12 voor het harvest- en opslagproces

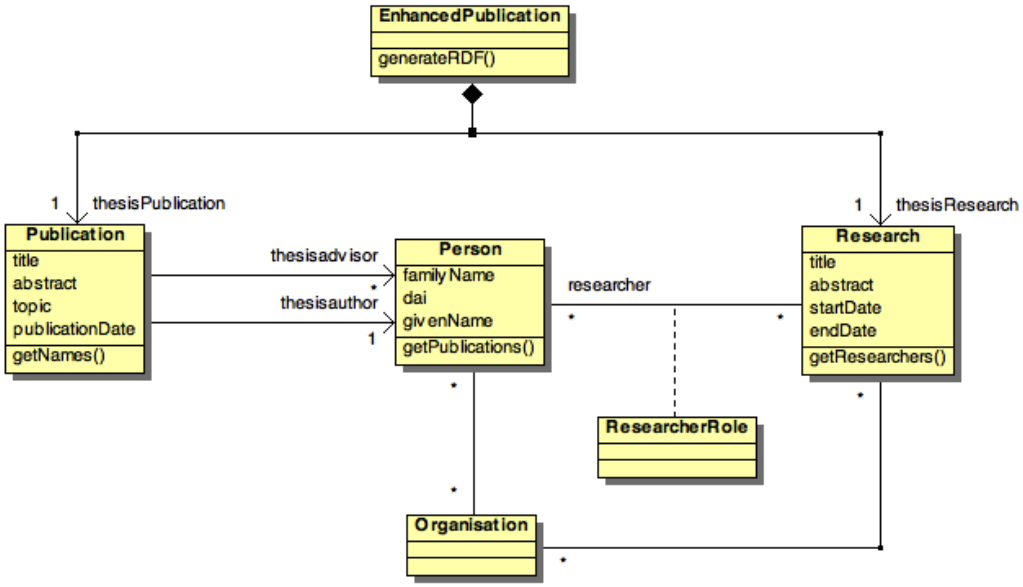
6.1.2 Iteratie demonstrator II: onderzoek aan publicatie koppelen

Demonstrator II had een eigen iteratie binnen de elaboration fase. Hierin is onderzocht of de combinatie van NOD-gegevens, researchers, onderzoeksorganisaties etc, en publicaties vanuit de verschillende repositories voldoende aanknopingspunten biedt om eenduidig vast te stellen of een publicatie de uitkomst is van een onderzoek.

Om de kans op een match te vergroten, heb ik besloten voor deze demonstrator me te beperken tot de promotie-onderzoeken. Bij een promotie-onderzoek worden de promotoren en de promovendus in de NOD bijgehouden. Veel promotoren zijn voorzien van een DAI. Bij een publicatie in de repository wordt vermeld of het een doctoral thesis betreft. Ik heb doctoral theses waarbij de auteur en/of promotoren in de metadata van de publicatie geen DAI hadden, buiten beschouwing gelaten.



Figuur 16: PSD voor het identificeren van het thesis onderzoek bij een gegeven publicatie



Figuur 17: class diagram uit het analysis model voor genereren van enhanced publications

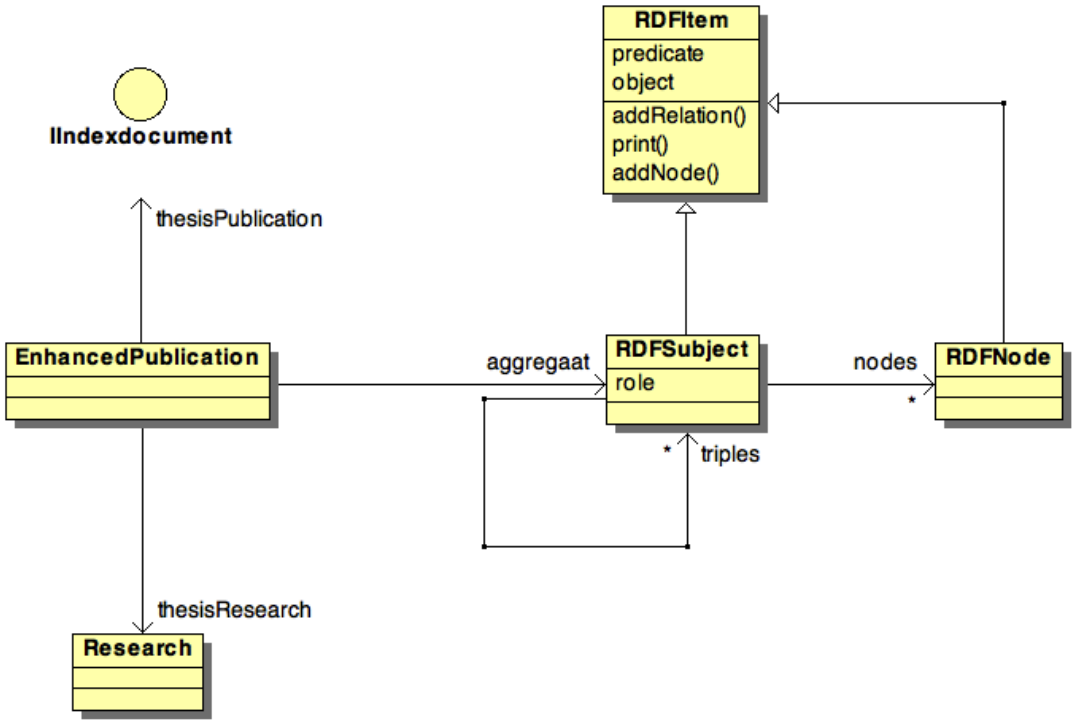
Na wat handwerk (dat voornamelijk bestond uit het toevoegen van de hieruit voortvloeiende DAI's in de NOD) kon vastgesteld worden dat er op deze manier zo'n twaalfhonderd koppelingen publicatie-onderzoek gemaakt kunnen worden.

Een EnhancedPublication is de combinatie van een Research met de daaruit voortvloeiende Publication. Dit zijn dezelfde classen als bij demonstrator I. De AuthorRole is hier weergegeven als twee associaties tussen Publication en Person aangezien niet alle MARC-codes van belang zijn; alleen de author en thesis-advisor worden gebruikt.

Vervolgens moest de vraag beantwoord worden hoe deze koppeling te visualiseren: of in de bestaande structuur van de NARCIS onderzoekspagina's, of in het kader van het nieuwe Enhanced Publication Project. Dit laatste kan waarschijnlijk op meer support rekenen, aangezien dit het tender-thema van SURFfoundation voor 2011 is. Daarom hebben we dan ook besloten demonstrator II onder te brengen in het Enhanced Publication Project.

Het format van de koppeling moet dan RDF[16] zijn, met de ontologiën zoals die binnen het SURF-project zijn aangewezen. In deze fase heb ik ervoor gekozen om slechts een globale opzet te maken van de data en de manier waarop die in RDF gegoten kan worden. De precieze invulling van de RDF heb ik vrijgelaten om in de construction fase nog verder uit te kunnen werken. Een eerste prototype is werkend gemaakt en het resultaat is getoond aan de projectleider van het Enhanced Publication Project. De tool die hiervoor gebruikt is, is de InContext visualiser van Q42 die binnen het SURF-project is ontwikkeld. Dit is een dynamische visualisatie waarmee door de Enhanced Publication genavigeerd kan worden.

16 Resource Description Framework, zie 10.10



Figuur 18: De RDF triples waar een Enhanced Publication uit opgebouwd is. Het IIndexdocument en het Research zijn bestaande classen.

```
<rdf:RDF [...] >
  <rdf:Description rdf:about="oai:dspace.library.uu.nl:1874/14036">
    <ore:describes rdf:resource="http://narcis.nl/[...]/oai:dspace.library.uu.nl:1874/14036"/>
    <dcterms:creator rdf:resource="http://www.narcis.nl"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://narcis.nl/[...]/oai:dspace.library.uu.nl:1874/14036">
    <rdf:type rdf:resource="http://www.openarchives.org/ore/terms/Aggregation" />
    <ore:aggregates rdf:resource="oai:dspace.library.uu.nl:1874/14036" />
    <ore:aggregates rdf:resource="OND1292696" />
    <dcterms:title>Doctoral Thesis: The Development of Personality [...]</dcterms:title>
  </rdf:Description>
  <escape-pubtypes:DoctoralThesis rdf:about="oai:dspace.library.uu.nl:1874/14036">
    <dcterms:title>The Development of Personality and Problem Behaviour [...]</dcterms:title>
    <escape-project:isOutcomeDocumentOf rdf:resource="OND1292696" />
    <dcterms:description>This dissertation focuses on [...]</dcterms:description>
  </escape-pubtypes:DoctoralThesis>
  <foaf:Project rdf:about="OND1292696">
    <escape-system:resourceUri rdf:resource="http://narcis.nl/[...]/OND1292696" />
    <escape-project:endDate>11 / 2006</escape-project:endDate>
    <dcterms:title>Probleemgedrag en persoonlijkheid van jongeren</dcterms:title>
    <escape-project:startDate>05 / 2002</escape-project:startDate>
  </foaf:Project>
</rdf:RDF>
```

Citaat 4: RDF representatie van de aggregatie van een onderzoek en een publicatie

Citaat 4 toont de RDF waarin de relaties tussen de objecten gelegd zijn. Figuur 19 toont de minimale aggregatie van een Research Project en de daaruit voortgekomen Publication. Dit prototype is later in de construction fase verder uitgebouwd en aangevuld met andere relaties.

Figuur 19: InContext visualiser met de relatie tussen een research project en een publication

6.2 Construction fase

6.2.1 Iteratie demonstrator I: coworkers

De eerste iteratie had als doel de auteurs van de publicaties te identificeren en in het moowheel te tonen. De plaatsing van de coworkers langs de rand van het wheel had als voordeel dat er geen 'belangrijkste' coworker is; analoog aan een ronde tafel. De volgorde heeft natuurlijk wel gevolgen voor de lengte van de getekende relaties onderling en dus voor het overzicht van de visualisatie. Ik heb geëxperimenteerd met verschillende plaatsingen, maar heb voor dit probleem nog geen oplossing gevonden. Bij elke keuze van de plaatsing langs de rand is een worst case situatie te bedenken en omgekeerd.

De tweede iteratie hield zich bezig met de onderzoeken. Het wheel is uitgebreid met coworkers die bij de onderzoeken van de persoon betrokken waren. Bij de evaluatie is vastgesteld dat de data in de NOD hiervoor niet bruikbaar is, want vaak worden ook de projectleider en de secretaris van het onderzoek genoemd, terwijl die niet feitelijk bij het onderzoek betrokken zijn. Andersom komt het ook vaak voor dat de onderzoekers die het daadwerkelijke onderzoek uitvoeren niet genoemd worden. Dit heeft te maken met de bekostiging: bij het indienen van een onderzoeksvoorstel bij NWO is het van strategisch belang een grote naam als projectleider te hebben. De overige medewerkers zijn hiervoor minder belangrijk. Aangezien NARCIS zijn data van NWO[17] krijgt, klopt dit dus vaak niet

17 De Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) financiert duizenden toponderzoekers aan universiteiten en instituten via subsidies en onderzoeksprogramma's

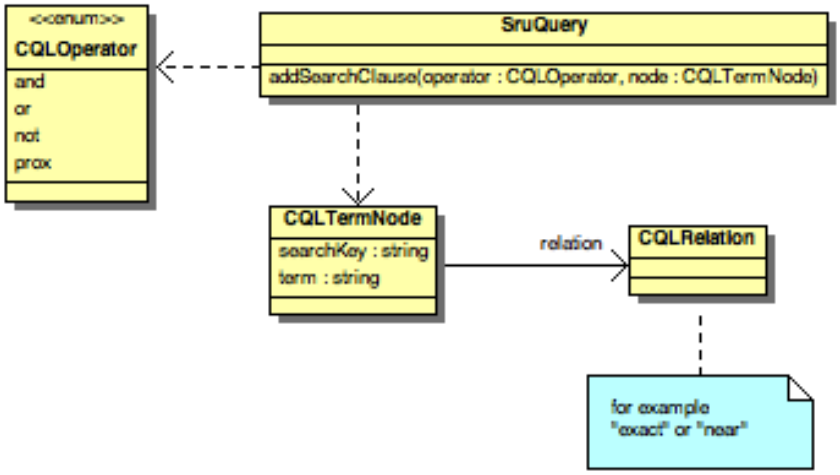
met de praktijk. Dit levert onverwachte en ongewenste samenwerkingen op.

6.2.2 Iteratie demonstrator II: onderzoek aan publicatie koppelen

In de construction fase is het prototype zoals dat bij het einde van de elaboration is opgeleverd omgebouwd tot volwaardig subsysteem. De visualisatie is niet meer gewijzigd

Om het eventueel opnemen in de NARCIS codebase te vergemakkelijken, heb ik er zoals gezegd voor gekozen zoveel mogelijk gebruik te maken van bestaande code. Maar tijdens het ontwikkelen wilde ik niet steeds de hele codebase gebruiken. Dus moesten er delen uit de bestaande pakketten 'losgetrokken' worden. Bijvoorbeeld het zoek-package was een integraal onderdeel van de NARCIS webportal. Om dit te kunnen gebruiken, was het noodzakelijk dat het een aparte library zou vormen.

Dit bracht nog een onverwacht probleem aan het licht. Tot dan toe werd een zoekquery opgebouwd door achtereenvolgens zoekclauses aan de query toe te voegen met een *addSearchClause* methode.



Figuur 20: beperkte weergave van de SruQuery uit het search-package

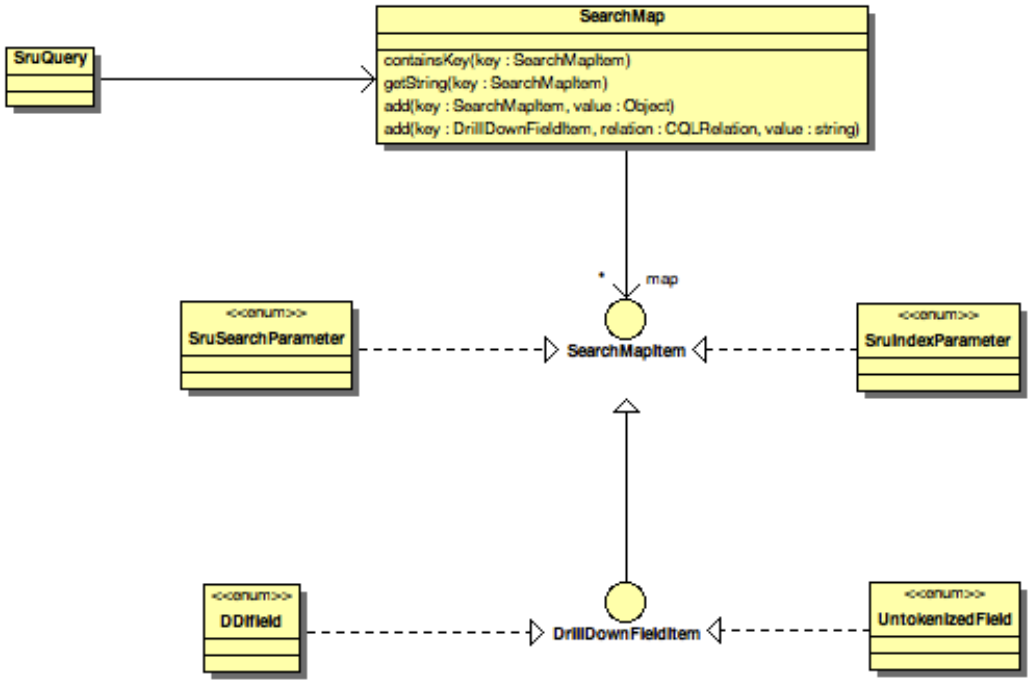
Dit werd achter de schermen steeds meteen omgezet naar een zoekstring, wat betekent dat de volgorde waarin de searchclauses toegevoegd werden, invloed had op de uiteindelijke zoekvraag. Dit levert onverwacht gedrag op en is daarom onwenselijk. Bovendien is de controle op de searchKeys (in de CQLTermNode zijn dat Strings) niet geregeld, waardoor het mogelijk is keys op te geven die in de NARCIS Index niet bestaan.

Om deze problemen te verhelpen, heb ik de functionaliteit die deze method leverde, verpakt in een SearchMap. Deze bevat een map: de keys komen uit een vooraf gedefinieerde set enums, de values zijn Objects die uiteindelijk als String in de zoekvraag terecht komen. Op deze manier kan elke key in het eindresultaat maar 1 keer voorkomen, en is gegarandeerd dat de key ook daadwerkelijk in de NARCIS Index bestaat. Om onderscheid te kunnen maken tussen de verschillende searchkeys heb ik ze onderverdeeld naar hun rol.

- SruSearchParameters zijn beschreven in het SRU protocol; deze staan vast.
- De SruIndexParameters zijn de 'kolomnamen' die in de NARCIS Index staan, deze

moeten dus overeenkomen.

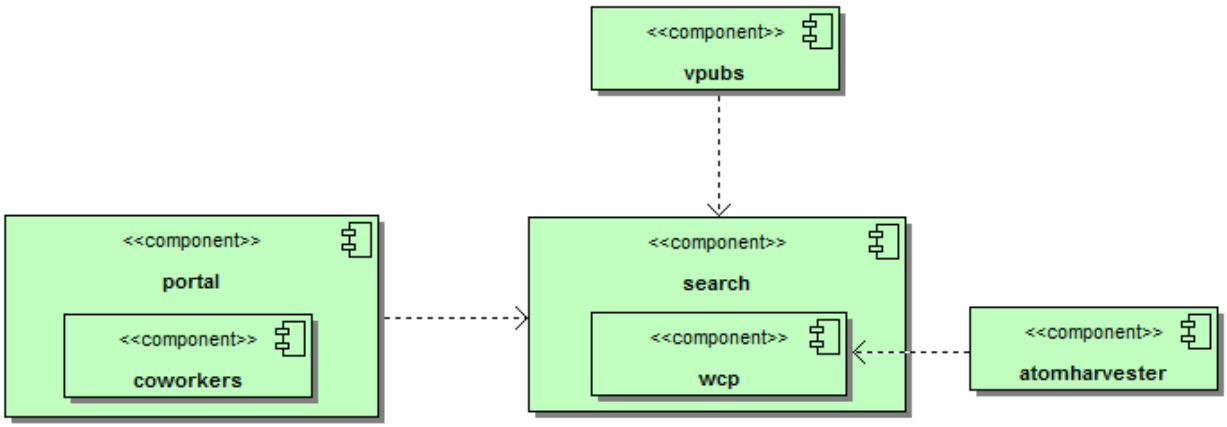
- Dat geldt ook voor de DrillDownItems.
- De set DDIFields is een echte subset van de UntokenizedFields. Bij de implementatie is deze restrictie verloren gegaan, aangezien ik met Java Enums heb gewerkt.



Figuur 21: De SruQuery kan niet meer handmatig aangepast worden, maar alleen via de gereguleerde SearchMap

Het TDD principe was bij deze refactoring zeer waardevol. Op het moment van aanvang werkte de zoekfunctionaliteit naar behoren, maar natuurlijk met de genoemde gebreken. Ik heb eerst testcases geschreven die in de toenmalige gang van zaken allemaal valideerden. Daarna heb ik een aantal testcases geschreven die de genoemde gebreken zichtbaar maakten. Dit gaf een solide basis om aan het refactoren te beginnen. Met de testcases in de hand, kon de progressie inzichtelijk gemaakt worden.

Uiteindelijk resulteerde dit in een nieuwe search-component, die zelfstandig te gebruiken is door bijvoorbeeld het vpubs-component, zoals te zien in figuur 22.



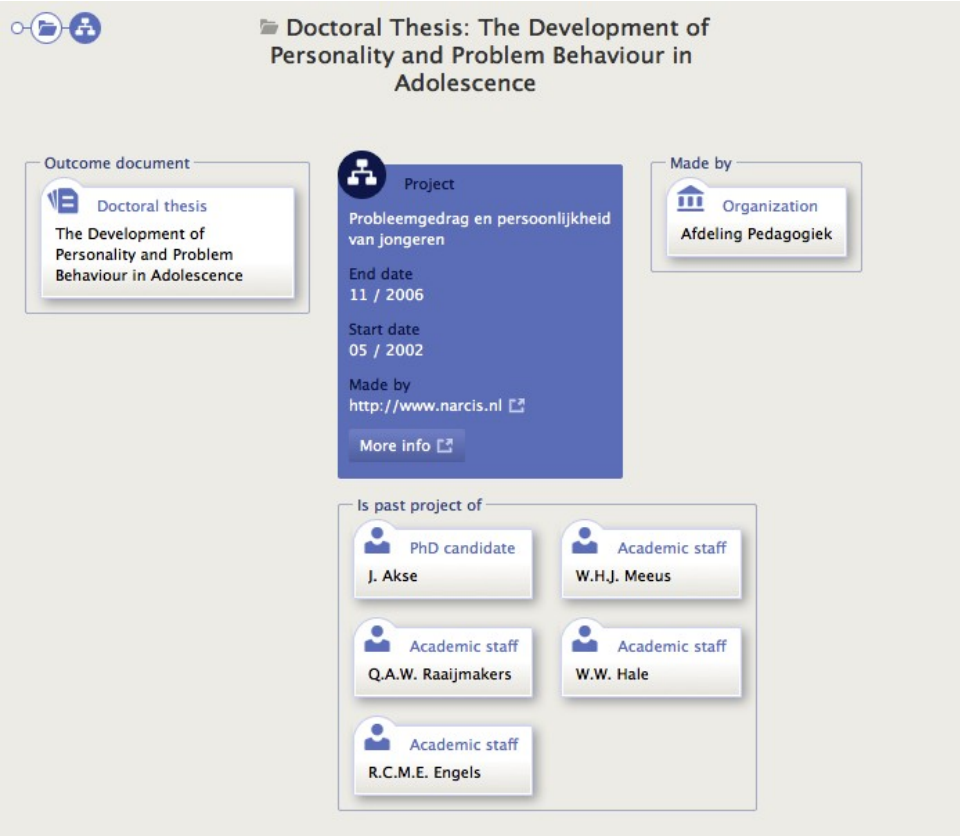
Figuur 22: component diagram

Tijdens het construeren van de Resource Maps (ReMs) kwamen er wat onvolkomenheden in het Enhanced Publication datamodel aan het licht. Daarom heb ik een meeting aangevraagd waarbij ook SURFfoundation aanwezig was. Dit heeft op 26 april plaatsgevonden. Een van de zaken die niet goed geregeld waren, was de status van de DAI. Het predicaat dai:dainr werd genoemd als een onderdeel van een bestaande ontologie, hoewel die nog niet bestaat. Ook de InContext visualiser voldeed niet aan de (impliciete) requirements. SURF heeft toegezegd deze zaken aan te kaarten bij de betreffende organisaties.

Ik heb drie iteraties gepland.

1	Prototype consolideren, automatisch genereren
2	ontologiën bestuderen en meer relaties toevoegen, beslissen waar de aggregatie stopt
3	Attributen van bestaande relaties toevoegen, vaststellen welke data geen tegenhanger in dit model heeft

Figuur 23 toont de aggregatie uitgaande van hetzelfde research – publication paar als figuur 19, maar dan aan het eind van de desbetreffende iteratie van de construction fase. Niet alleen zijn organisaties toegevoegd, maar ook de betrokken Personen, en de relaties onderling zijn gelegd. Bovendien is een dataset toegevoegd die ook bij de thesis betrokken was. Zie bijlage 10.14 voor de achterliggende RDF.



Figuur 23: InContext visualisatie van een Enhanced Publication. Centraal staat hier het Promotie Project met de bijbehorende Doctoral Thesis

6.3 Transition fase

In de transition fase heb ik gekeken naar de stappen die gezet moeten worden om de demonstrators in productie te kunnen nemen.

6.3.1 Iteratie demonstrator I

Demonstrator I is ver genoeg ontwikkeld en getest om in productie genomen te worden. De vraag die beantwoord moet worden is of de NARCIS projectleiding het aandurft de genoemde incompleetheid van de data te etaleren. Wanneer een onderzoeker naar zijn/haar coworker wheel kijkt, zal hij/zij daar namen missen. Dit ontbreken van DAI's was exact wat de demonstrator moest aantonen, maar vormt wellicht tegelijkertijd een struikelblok voor de in-productie-name van het wheel.

Een ander issue is de performance. Requirement 6 sprak over deze niet-functionele eis. Deze requirement was niet gekwantificeerd. Voor de doeleinden van de demonstrator was deze eis niet noodzakelijk. Ik heb er dan ook geen rekening mee gehouden bij het ontwikkelen. Wanneer het wheel in productie genomen gaat worden, zal dit wel opgelost moeten worden. De bottleneck is het opvragen van alle publicaties van een persoon uit de Index. Een mogelijke oplossing is om dit niet on-demand te construeren, maar asynchroon.

6.3.2 Iteratie demonstrator II

Om demonstrator II daadwerkelijk in productie te nemen, is het aanleveren van de ReMs niet voldoende. Ze moeten ook geharvest kunnen worden en de juiste velden moeten geïndexeerd worden om ze vindbaar te maken.

In de transition fase is een manier bedacht om dit te doen.

Er zijn verschillende opties, waarvan de meest voor de hand liggende OAI-PMH[18] en ATOM[19] zijn. OAI-PMH is wellicht overkill, ATOM is eenvoudiger te realiseren. NARCIS ondersteunt op dit moment alleen het OAI-PMH protocol. Om ATOM te kunnen harvesten moet dus een AtomHarvester ontwikkeld worden.

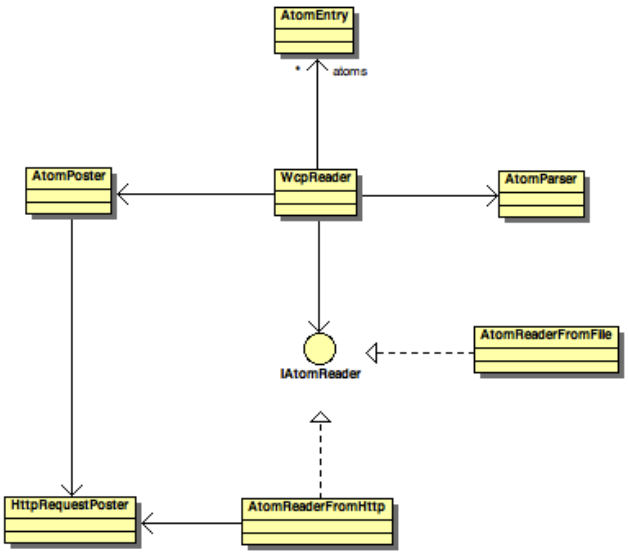
Een prototype voor de AtomHarvester heb ik ontwikkeld en gecodeerd. (zie figuur 24 & Error: Reference source not found) De Enhanced Publications moeten dus in ATOM formaat aangeboden worden. Dit is in eerste instantie zeer rechttoe rechtaan. Om het echt in productie te nemen, moet een aantal problemen eerst geadresseerd worden:

- Op dit moment wordt de relatie tussen een onderzoek en een publicatie nergens expliciet opgeslagen. Dat betekent dat om de Enhanced Publications te maken, alle thesis publicaties langs gelopen moeten worden, een kostbare bezigheid. Echter, de relatie expliciet vastleggen levert redundantie op, met alle synchronisatie problemen van dien.
- Op welke moment, if any, moet een Enhanced Publication aangepast worden? De huidige opbouw van de NOD is puur gericht op lopend onderzoek. De structuur is niet gebouwd op wijzigingen. Instituten worden opgeheven, gesplitst of fuseren,

18 Bijlage 10.12
19 Bijlage 10.13

deze historie is niet bij te houden in de huidige database. Ook wanneer een onderzoeker een instituut verlaat, wordt die relatie uit de NOD verwijderd. Het opnieuw genereren van Enhanced Publications heeft dus risico's.

- Heeft zo'n Enhanced Publication een eigen DOI? En moet die dus ook duurzaam aangeboden worden? De verwijzingen in de Enhanced Publication naar NARCIS objectpagina's kunnen verlopen, zie het vorige punt. Pas wanneer deze duurzaam zijn, kan de Enhanced Publication duurzaam geresolved worden.



Figuur 24: AtomHarvester, gerealiseerd als WcpReader

7 Evaluatie

In dit hoofdstuk evalueer ik eerst het proces dat ik heb gevolgd gedurende deze afstudeeropdracht. In het tweede deel heb ik de evaluatie van de producten weergegeven. Deze evaluatie heb ik deels gezamenlijk met mijn bedrijfsmentor uitgevoerd.

7.1 Evaluatie van het proces

In de evaluatie van het proces heb ik gekeken naar hoe de gemaakte keuzes uiteindelijk hebben uitgepakt. Hierbij doe ik aanbevelingen ter verbetering en geef leerpunten voor mezelf aan.

7.1.1 Unified Process als ontwikkelmethode

Ik heb indertijd gekozen voor UP omdat ik daar een duidelijke scheiding zag tussen een onderzoeksfase en een ontwikkelfase. Dit heeft in de praktijk ook zo gewerkt. Tijdens het onderzoek heb ik me niet beziggehouden met realisatie problemen. Zo kon bijvoorbeeld tijdens de brainstormsessie vrijuit geassocieerd worden en werden de deelnemers niet gehinderd door veronderstelde technische beperkingen.

De beide demonstrators putten uit hetzelfde domein. Het leek mij daarom wenselijk ze in één project te ontwikkelen. Dit heeft soms vruchten afgeworpen, soms ook niet zoals uit de volgende paragrafen blijkt.

7.1.2 Inception fase

Ik had verwacht dat de gesprekken met de stakeholders op een meta-niveau gevoerd konden worden. Maar stakeholders bleken beter te kunnen weergeven welke oplossingen zij zien dan aan te kunnen geven wat de achterliggende behoeften zijn. Tijdens de eerste gespreksronde heb ik het gesprek laten gaan over deze oplossingen en mogelijke toepassingen. Vervolgens heb ik zelf daaruit achterliggende criteria opgesteld.

Daarna wilde ik een aantal alternatieven beschrijven en hieruit samen met de stakeholders de 'beste' kiezen. Ik wilde een ordinale methode zoals de datum-methode van Pugh [M&T05] toepassen[20]. Dit is een recursieve methode, waarbij de output uit de ene ronde de input voor de volgende vormt. Ik had deze methode graag toegepast, omdat ik verwacht dat er op deze manier een convergerende discussie ontstaat, het inzicht in de criteria wordt vergroot en er mogelijkerwijs ook nieuwe alternatieven worden bedacht. Wegens het herhalende, convergerende karakter van deze methode is het noodzakelijk dat de betrokkenen met elkaar in gesprek zijn en blijven.
Het was echter niet mogelijk alle stakeholders op hetzelfde moment en voor langere tijd bij elkaar te krijgen. Daarom heb ik uiteindelijk een andere, kardinale methode gebruikt, zie paragraaf 5.2. Hierbij is het niet noodzakelijk dat de stakeholders met elkaar in gesprek gaan over de alternatieven.

Ik heb de tweede gespreksronde gebruikt om de opgestelde criteria bij de stakeholders te toetsen en te vergelijken. De gekozen methode heeft als voordeel dat degene die beslist over de criteria een heldere keuze voorgelegd krijgt: welk van deze twee criteria vindt de

20 Zie bijlage 10.11 Datum-methode van Pugh

beslisser het belangrijkst? De stakeholder hoeft dus niet de alternatieven met elkaar te vergelijken, een veel lastiger probleem.

Voor deze methode is het nodig dat de alternatieven op alle criteria gescoord worden. Dit is geen objectieve score, hoewel dat door de getalletjes wel zo lijkt. Hierover kan eindeloos gediscussieerd worden, zonder dat dit tot meer draagvlak of overeenstemming leidt. Ik heb de discussie gestopt nadat ik had vastgesteld dat elk lid van het NARCIS ontwikkelteam het met de strekking van de getallen eens was. De precieze grootte van de getallen bleek een minder belangrijk punt te zijn. Een ordinale methode was beter geweest omdat de discussie dan was gegaan over de alternatieven in plaats van de scores.

7.1.3 Elaboration fase

Omdat de twee demonstrators betrekking hebben op hetzelfde domein, heb ik ze tegelijkertijd in een project ontwikkeld. Eén project voor de beide demonstrators klonk vooraf als een haalbare en nuttige aanpak. In de praktijk zit er immers een overlap tussen de modellen. Het gezamenlijk iteratief ontwerpen heeft wel wat vruchten (tijdswinst, analyse demonstrator II is een voortzetting van demonstrator I) afgeworpen, maar niet in die mate die ik had verwacht. Hoewel ze hetzelfde domein beschrijven, ligt de nadruk van beide zodanig anders dat de overlap niet groot is.

7.1.4 Construction fase

Bij de construction fase, nadat de architectuur vastgesteld was, lag de focus bij de twee demonstrators zo verschillend, dat het voordeel van het samen implementeren niet werd behaald.

Bij demonstrator I was het inpassen in de NARCIS Portal een belangrijk aandachtspunt. De nadruk lag op de visualisatie. De achterliggende relaties waren niet complex. Bij demonstrator II daarentegen ging het juist om die relaties. Er werd een nieuwe relatie gelegd tussen een onderzoek en een publicatie. Van de relaties daaromheen heb ik bepaald welke in de Enhanced Publication opgenomen moesten worden. De visualisatie van demonstrator II was snel bepaald, namelijk de InContext visualisatie van SURF.

Tijdens het ontwikkelen van de prototypes en de daaruit volgende demonstrators heb ik volgens TDD gewerkt, dus op een Agile manier. Bij het maken van de prototypes in de elaboration fase heb ik testen geschreven om de methodes te ontwikkelen. Dit heeft bijgedragen aan een architectuur met losse koppeling, aangezien dat de enige manier is om op deze manier te unittesten. Nadat de architectuur was vastgesteld, heb ik TDD ingezet om de demonstrators verder te ontwikkelen. De integratie-testen zijn ook op deze manier uitgevoerd.

Ook bij het refactoren van de search-component heb ik gebruik gemaakt van de lessen van TDD. De officiële weg is om eerst de test te schrijven, te falen, de code te schrijven zodat de test slaagt en dan de code te refactoren. Tijdens het refactoren van de search-component heb ik het dus iets anders gedaan, de functionaliteit was er al, maar nog niet in testen beschreven. Ik heb de testen opgesteld, deze slaagden, en daarna de component gerefactored zodanig dat de testen nog steeds slagen. Bij het constateren van een tekortkoming heb ik testgevallen geconstrueerd die het probleem aantonen, en daarna de code gewijzigd, zie paragraaf 6.2.2. Deze manier van werken zorgde ervoor dat de bestaande functionaliteit behouden is, de code beter testbaar is en de component ook

buiten de Portal herbruikbaar is.

7.2 Evaluatie van de producten

7.2.1 Onderzoeksrapport

Het uitvoeren van het onderzoek en het schrijven van het rapport heeft de inception fase structuur gegeven. Doordat de criteria van de stakeholders vroeg in het proces bekend en helder omschreven waren, was verdere input van de stakeholders wenselijk maar niet noodzakelijk. Dit bevorderde de doorloop. De criteria gaven bovendien duidelijk richting aan de brainstorm naar alternatieven.

Het rapport heeft geleid tot een duidelijk resultaat. De twee gekozen demonstrators scoorden bij alle stakeholders hoog. Dit gaf een solide basis voor het vervolg. Bovendien was daardoor een belangrijk deel van de requirements al bepaald.

Persoonlijk vond ik het interessant om te zien dat A4 (*profielpagina's koppelen, zoals LinkedIn, Facebook*) en A2 (*extra bronnen raadplegen, zoals PubMed of Elsevier*) zo laag scoren, terwijl ze door alle stakeholders genoemd zijn als iets wat ze graag zouden willen. Bij nader inzien blijkt dat de kans op onjuiste data, en privacy overwegingen daarbij gemakkelijk genegeerd worden. Doordat bij dit onderzoek alle alternatieven op dezelfde criteria gescoord zijn, komen deze 'tekortkomingen' wel boven water.

7.2.2 Demonstrator I

Het coworkerwheel toont heel duidelijk voor een onderzoeker welke DAI's uit zijn omgeving ontbreken. Hij kan dus actief vragen binnen zijn instituut om deze metadata aan de publicaties toe te voegen. Dit geeft een duidelijke incentive. Het heeft dus aan zijn doel beantwoord.

7.2.3 Demonstrator II

Hoewel de beide iteraties binnen de construction fase redelijk ver uit elkaar lagen, was het toch nuttig de beide demonstrators in één project te ontwikkelen. Samen geven ze een goed beeld van wat mogelijk is met de DAI. Demonstrator I laat op een nieuwe manier bestaande relaties zien, en biedt hierdoor nieuwe inzichten. Demonstrator II daarentegen legt een nieuwe relatie (die in de werkelijkheid wel bestaat, maar nog niet in de databanken) en verbindt zo twee objecten uit verschillende typen bronnen.

Bij het ontwikkelen van demonstrator II heb ik de volgende aanname gedaan [TJR01]:

“Wanneer de DAI van de promovendus van het onderzoek gelijk is aan de DAI van de auteur van het proefschrift, kan aangenomen worden dat dat onderzoek bij die publicatie hoort.”

Deze aanname blijkt niet altijd waar te zijn, soms is een onderzoeker meer dan één keer gepromoveerd. De demonstrator maakt hier geen onderscheid tussen, en voegt bij alle publicaties het eerst gevonden onderzoek toe. Dit is onderzocht en geconstateerd tijdens de elaboration fase. Met toestemming bedrijfsmentor heb ik besloten de Enhanced Publication wel op deze manier te genereren. Onderzoekers met meer dan één promotie-onderzoek worden voorlopig genegeerd. Later kunnen uitgebreidere vergelijkingen toegepast worden, bijvoorbeeld op titel, promotor, universiteit etc.

8 Beroepstaken

In het afstudeerplan, [TJR07], zijn de volgende competenties/beroepstaken genoemd [AICTM01]:

- **1.2 Voorbereiden en opstarten softwareontwikkeltraject**
- **3.2 Ontwerpen systeemdeel**
- **3.3 Bouwen applicatie**
- **3.5 Uitvoeren van en rapporteren over het testproces**

In dit hoofdstuk worden deze beroepstaken langsgelopen en wordt telkens aangegeven hoe en waar die beroepstaak is behaald. De letterlijke tekst uit de beroepstaak is cursief weergegeven, de invulling, waar en hoe, staat ingesprongen.

Ad 1.2 Voorbereiden en opstarten softwareontwikkeltraject

Op hoofdlijnen analyseren en beschrijven van een vraagstuk of probleem in de informatievoorziening, voortkomend uit een bedrijfskundig vraagstuk of uit niet-administratieve toepassingen.

> Deze analyse is verwoord in het Plan van Aanpak [TJR04], met name in de Projectdefinitie en de Projectaanpak.

Maken van een omgevingsanalyse en in kaart brengen van een IV en/of IT landschap.

> Het resultaat van deze omgevingsanalyse en het onderzoeken van mogelijke toepassingen is beschreven in het onderzoeksrapport [TJR01].

Ad 3.2 Ontwerpen systeemdeel

Beschrijven van systeemdelen (subsystemen, componenten, modules), hun onderliggende structuur en het gedrag in detail, zodanig dat bouwen van het systeemdeel mogelijk is.

> De structuur en het gedrag van de demonstrators is ontworpen door middel van een analyse en een design model. Aan de hand daarvan zijn de demonstrators daadwerkelijk gebouwd.

Ontwerpen (grafische) user interface.

> Bij de keuze van de demonstrators zoals beschreven in het onderzoeksrapport [TJR01] en bij de keuze van een geschikte visualisatie voor demonstrator I [TJR03] is nadrukkelijk aandacht besteed aan de user interface zodat ze inderdaad als *demonstrator* gebruikt kunnen worden.

Ad 3.3 Bouwen applicatie

Verfijnen en/of transformeren van het (detail)ontwerp van de systeemdelen (subsystemen, componenten, modules) van een applicatie.

Uitgaande van het design-model heb ik gezocht naar bruikbare functionaliteit binnen de met Java meegeleverde libraries en de facto standaarden als apache-commons.

Bouwen en documenteren van de systeemdelen.

Deze libraries heb ik toegepast bij het bouwen van de applicatie in de IDE Eclipse. Deze applicatie is gedocumenteerd met behulp van Javadoc. Bovendien is de functionaliteit verwoord in de requirements en gebruikt bij de unit testen, zoals bij TDD noodzakelijk. Ontwikkelen en Testen hebben plaatsgevonden binnen de daartoe ingerichte omgevingen van de OTAP-straat. Vanzelfsprekend is hierbij gebruik gemaakt van het bij de afdeling *Onderzoek Informatie* gebruikte version control system Subversion.
Bij het bouwen van het systeemdeel is de bestaande architectuur beter herbruikbaar gemaakt door het refactoren van de search-component.

Systeemdelen samenstellen tot een werkende applicatie.

De demonstrators zijn ingepast in de bestaande architectuur. Demonstrator I *coworkers tonen* is beschikbaar binnen de NARCIS Portal. Van demonstrator II *onderzoek aan publicatie koppelen* is in de transition fase onderzocht welke stappen genomen moeten worden om dit op te nemen in de huidige infrastructuur. De AtomHarvester is in prototype ontwikkeld. In het bedrijfsproces moeten beslissingen genomen worden door de NARCIS projectleiding met betrekking tot de toekomst van NARCIS. Mocht besloten worden om hiermee door te gaan, dan staat alles in de steigers om verder uitgebouwd te worden.

Ad 3.5 Uitvoeren van en rapporteren over het testproces

Opstellen testscript, waaronder het specificeren van de testgegevens, de benodigde startsituatie en de uitvoerverwachtingen per testgebied. Uitvoeren testplan en opstellen rapportage.

Door tijdens de Elaboration en Construction fase te werken volgens Test Driven Development is gestuurd op het voldoen aan de requirements. TDD zorgt ervoor dat er altijd een werkend stuk code ligt, dat in ieder geval een deel van de gevraagde functionaliteit levert.

Door het uitvoeren van unittesten, integratie-testen en functionele acceptatietesten (FAT) is de kwaliteit van het product vastgesteld. Hierbij zijn testcases geconstrueerd, startsituatie beschreven in XML documenten, testscripts gemaakt en uitgevoerd. De FAT is door de bedrijfsmentor uitgevoerd.

9 Literatuur en websites

Literatuur

TJR01: T.J. Reijnhoudt, onderzoeksrapport naar het tonen van de meerwaarde van de DAI, 2011
M&T01: Jim Arlow; Ila Neustadt, UML 2 and the Unified Process, 2005
TJR05: T.J. Reijnhoudt, Glossary, 2011
TJR03: T.J. Reijnhoudt, Uitwerking demonstrator 'Coworker' , 2011
TJR06: T.J. Reijnhoudt, Uitwerking demonstrator 'Onderzoek aan Publicatie koppelen', 2011
M&T03: Kent Beck, Test Driven Development by Example, 2003
M&T06: Martin Fowler, Refactoring - Impoving the Design of Existing Code, 1999
TJR02: T.J. Reijnhoudt, Bepalen van weegfactoren, 2011
TJR08: T.J. Reijnhoudt, Presentatie coworker visualisatie, 2011
M&T05: S. Pugh, Concept selection: a method that works, 1981
TJR07: T.J. Reijnhoudt, Afstudeerplan 2011-1.1, 2011
AICTM01: , Beroepstaken van de Opleiding Informatica, 2009
TJR04: T.J. Reijnhoudt, Plan van Aanpak, 2011

Websites

http://dans.knaw.nl

http://www.knaw.nl/Pages/DEF/26/157.bGFuZz1OTA.html

http://www.narcis.nl/person/info:eu-repo/dai/nl/304357960

http://www.loc.gov/loc.terms/relators/

http://wiki.surffoundation.nl/display/standards/Use+of+MODS

http://www.nwo.nl/

http://wiki.surffoundation.nl/display/vp/Home

http://www.loc.gov/standards/sru/

http://xmlns.com/foaf/spec/

http://www.w3.org/RDF/

http://www.openarchives.org/OAI/openarchivesprotocol.html

http://www.ietf.org/rfc/rfc5005.txt

10 Bijlagen

De volgende bijlagen zijn opgenomen in extern document.

10.1 Afstudeerplan

10.2 Afkortingenlijst

10.3 Plan van Aanpak

10.4 Gespreksverslagen stakeholders

10.5 Onderzoeksrapport

10.6 Weegfactoren

10.7 Uitwerking demonstrator Coworker

10.8 Presentatie Coworker visualisatie

10.9 Uitwerking demonstrator Onderzoek aan Publicatie koppelen

10.10 RDF & ontologiën

10.11 Datum-methode van Pugh

10.12 Harvesten

10.13 ATOM

10.14 Volledig RDF voorbeeld