

Afstudeerplan

Informatie afstudeerder en gastbedrijf

Afstudeerblok: 2011-1.1 (start uiterlijk 7 februari 2011)
Startdatum uitvoering afstudeeropdracht: 7 februari 2011
Einddatum uitvoering afstudeeropdracht: 3 juni 2011
Inleverdatum afstudeerdossier volgens jaarrooster: 3 juni 2011

Studentnummer: 10102515
Achternaam: mw Reijnhoudt
Voorletters: T.J.
Roepnaam: Linda
Adres: van Cittersstraat 28 b
Postcode: 3022 LK
Woonplaats: Rotterdam
Telefoonnummer:
Mobiel nummer: 06 337 55 240
Privé emailadres: linda@reijnhoudt.nl

Opleiding: Informatica
Locatie: Den Haag
Variant: deeltijd

Naam studieloopbaanbegeleider: Helene Weenink
Naam begeleider/examinator: Tim Cocx
Naam expert/examinator: Gert Jan Timmerman

Naam bedrijf: Koninklijke Nederlandse Akademie der Wetenschappen
Afdeling bedrijf: OnderzoekInformatie
Bezoekadres bedrijf: Kloveniersburgwal 29
Postcode bezoekadres:
Postbusnummer:
Postcode postbusnummer:
Plaats: Amsterdam
Telefoon bedrijf:
Telefax bedrijf:
Internetsite bedrijf: www.knaw.nl

Achternaam opdrachtgever: dhr. Harmsen
Voorletters opdrachtgever: H.
Titulatuur opdrachtgever: Dr.
Functie opdrachtgever: Adjunct Directeur
Doorkiesnummer opdrachtgever: 06 30420605
Email opdrachtgever: henk.harmsen@dans.knaw.nl

Achternaam bedrijfsmentor: dhr Baars
Voorletters bedrijfsmentor: C.
Titulatuur bedrijfsmentor:
Functie bedrijfsmentor: Procescoördinator ICT
Doorkiesnummer bedrijfsmentor: 020 5510 986
Email bedrijfsmentor: chris.baars@bureau.knaw.nl

NB: bedrijfsmentor mag dezelfde zijn als de opdrachtgever

Doorkiesnummer afstudeerder: 020 5510 987
Functie afstudeerder (deeltijd/duaal): ontwikkelaar

Titel afstudeeropdracht:

De meerwaarde van de Digital Author Identification

Opdrachtomschrijving

1. Bedrijf

De Koninklijke Nederlandse Akademie der 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.

Binnen de KNAW houdt de afdeling OnderzoekInformatie (8 FTE) zich bezig met het toegankelijk en transparant maken van Nederlands publiek gefinancierd onderzoek en het zichtbaar maken van kennis van onderzoekers en onderzoeksinstituten. Een van de manieren waarop dat gebeurt, is door middel van het product Narcis.

Narcis harvest de repositories van alle Nederlandse universiteiten en een groot aantal onderzoeksinstituten. Bovendien heeft het toegang tot de Nederlandse Onderzoeks Databank (NOD), waar al het onderzoek waar NWO-geld naartoe gaat, is aangemeld.

2. Probleemstelling

Een aantal jaar geleden is de DAI ingevoerd, de Digital Author Identification, en langzaam maar zeker vindt die zijn weg in de repositories, waardoor het mogelijk is, ondanks verschillende schrijfwijzen van namen en voorletters, alle publicaties van een auteur te vinden.

Om te laten zien wat er gedaan kan worden als al deze data is gekoppeld, is er ook een portal gebouwd: www.narcis.nl. Hier wordt met behulp van een java-applicatie de data getoond. De portal maakt nog weinig gebruik van het feit dat de data van alle universiteiten en onderzoeken beschikbaar en koppelbaar is. Op dit moment kan op DAI-nummer een persoonspagina gemaakt worden, waarop de publicaties van die persoon getoond worden.

De afdeling OnderzoekInformatie wil graag dat meer instituten hun repositories voorzien van DAI data. Om dit te bereiken zal aangetoond moeten worden wat de meerwaarde van die effort is.

Ze wil laten onderzoeken welke informatie uit de data gehaald kan worden op basis van de DAI. Deze data bestaat uit zowel de geharveste publicaties en datasets en de gegevens uit de NOD. Bovendien wil ze een tweetal prototypes laten ontwikkelen om in het werkveld te tonen.

3. Doelstelling van de afstudeeropdracht

Door middel van een onderzoek naar de mogelijkheden bepalen welke prototypes gebouwd zullen worden. Ontwerpen, bouwen en testen van de prototypes.

4. Resultaat

De afdeling beschikt over een onderzoeksrapport waarin mogelijkheden worden geschetst om de DAI te 'verkopen' bij afnemers van de informatie. Bovendien beschikt de afdeling over twee prototypes waarmee zij aan de repository-beheerders kan laten zien wat de meerwaarde van het landelijk harvesten van met DAI verrijkte publicaties is. Deze prototypes worden zover ontwikkeld dat ze opgenomen kunnen worden in het portfolio van diensten en producten van de afdeling.

5. Uit te voeren werkzaamheden, inclusief een globale fasering, mijlpalen en bijbehorende activiteiten

nr	activiteit	aantal dagen	mijlpalen	beroepst aken
1	schrijven plan van aanpak	3	PvA	1.2
2a	<i>RUP inception fase</i> onderzoek naar mogelijkheden tot promoten van de DAI	10	onderzoeks- rapport	1.2
2b	in samenspraak met opdrachtgever twee ideeën uitkiezen		shortlist	
3	plannen ontwikkeltraject	2	globale planning met mijlpalen	1.2
4	<i>RUP Elaboration fase</i> vaststellen requirements / scenario opstellen testplan	2	requirements testplan	3.4
5	bekijken standaardoplossingen en bibliotheken	3		
6	<i>RUP Elaboration fase</i> ontwerpen prototype 1	4	ontwerp	3.2
7	<i>RUP Construction fase</i> ontwikkelen prototype 1 volgens TDD vaststellen kwaliteit prototype ¹ 1	15	code (java, html, javascript, css) testcode	3.3 1.4 3.5
8..11	stappen 4 t/m 7 voor prototype 2	24		
12	<i>RUP Transition</i> voorbereiden opname in portfolio van de afdeling	6	transitie- document	
13	samenstellen portfolio: schrijven afstudeerverslag	15		

Table 1: overzicht van de werkzaamheden

In de inception fase wordt een onderzoek uitgevoerd naar de mogelijkheden die de DAI biedt. Daarna wordt prototype 1 ontworpen en ontwikkeld, en in een volgende iteratie prototype 2. Hierna volgt een transition fase.

6. Op te leveren (tussen)producten

Zie de kolom 'mijlpalen' in het vorige onderdeel.

¹ hier opgenomen als aparte activiteit, is natuurlijk onderdeel van zowel ontwerpen als bouwen

7. Te demonstreren competenties en wijze waarop

Overzicht van de beroepstaken die in deze afstudeeropdracht aan bod komen, en de activiteiten waarin ze gedemonstreerd worden. Bij elke beroepstaak staat een toepasselijk deel geciteerd uit [BEROEPSTAKEN VAN DE OPLEIDING INFORMATICA Academie voor ICT & Media, juni 2009, versie 1.1].

1.2 Voorbereiden en opstarten softwareontwikkeltraject (niveau 2)

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

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

- onderzoek naar mogelijkheden tot promoten van de DAI
- schrijven plan van aanpak
- plannen ontwikkeltraject

3.2 Ontwerpen systeemdeel (niveau 3)

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

Ontwerpen (grafische) user interface.

- ontwerpen prototype

3.3 Bouwen applicatie (niveau 3)

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

Bouwen en documenteren van de systeemdelen.

Systeemdelen samenstellen tot een werkende applicatie.

- ontwikkelen prototype volgens TDD

3.5 Uitvoeren van en rapporteren over het testproces (niveau 3)

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

Uitvoeren testplan en opstellen rapportage.

- opstellen testplan
- ontwikkelen prototype volgens TDD
- vaststellen kwaliteit prototype

RDF & ontologiën

RDF is een generieke methode om feiten te noteren. Een feit bestaat uit drie delen: een subject, een predicaat en een object. Dit worden ook wel triples genoemd.¹

Er zijn verschillende formaten waarin een en ander gezet kan worden, zoals N3 of XML. De volgende voorbeelden zijn in N3.

```
@prefix : <http://www.example.org/> .
:john    a          :Person .
:john    :hasMother  :susan .
:john    :hasFather  :richard .
:richard :hasBrother :luke .
```

Citaat 1: vier RDF triples volgens de N3 notatie

Ditzelfde voorbeeld kan ook in XML genoteerd worden.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:ns="http://www.example.org/#">
  <ns:Person rdf:about="http://www.example.org/#john">
    <ns:hasMother rdf:resource="http://www.example.org/#susan" />
    <ns:hasFather>
      <rdf:Description rdf:about="http://www.example.org/#richard">
        <ns:hasBrother rdf:resource="http://www.example.org/#luke" />
      </rdf:Description>
    </ns:hasFather>
  </ns:Person>
</rdf:RDF>
```

Citaat 2: XML is veel meer verbose dan N3

Een ontologie beschrijft vervolgens de regels die op die feiten toegepast kunnen worden om daar mee te redeneren. Deze ontologiën zijn dus domeinspecifiek. In N3-notatie wordt de relatie 'hasAunt' op de volgende manier afgeleid van de relaties 'hasFather' en 'hasSister':

```
{ ?a :hasFather ?b . ?b :hasSister ?c . } => { ?a :hasAunt ?c } .
```

¹ <http://www.w3.org/RDF/>
<http://rdfabout.com/quickintro.xpd>

RDF in het Enhanced Publication Project

SURFfoundation heeft voor het Enhanced Publication project de volgende core ontologiën voorgesteld²:

Prefix	Namespace URI	Description
dc	http://purl.org/dc/terms/	Dublin Core terms
foaf	http://xmlns.com/foaf/0.1/	Friend of a Friend
dai	http://purl.org/info:eu-repo/dai#	Digital Author Identifier
eurepo	http://purl.org/info:eu-repo/semantics/	info:eu-repo Controlled Vocabulary for Repositories in the EU context
wdrs	http://www.w3.org/2007/05/powder-s#	Web Description Resources (Powder-S)
ore	http://www.openarchives.org/ore/terms/	ORE vocabulary terms
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF vocabulary terms

Daarnaast zijn ook de deze community ontologiën toegestaan:

Prefix	Namespace URI	Description
swan	http://swan.mindinformatics.org/ontologies/1.2/discourse-relationships/	Discourse Relationships
escape-projects	http://purl.utwente.nl/ns/escape-projects.owl	Research Project
escape-annotations	http://purl.utwente.nl/ns/escape-annotations.owl	Relation Annotation
escape-agents	http://purl.utwente.nl/ns/escape-agents.owl	Agents
escape-system	http://purl.utwente.nl/ns/escape-system.owl	System
escape-publicationtypes	http://purl.utwente.nl/ns/escape-pubtypes.owl	Publication Types

² <http://wiki.surfoundation.nl/display/vp/Resource+Maps+in+RDF+XML#ResourceMaps+in+RDF+XML-Ontologies+and+their+Namespaces>

Datum-methode van Pugh

De datum-methode van Pugh is een ordinale methode om alternatieven te rangschikken, waarbij de alternatieven per criterium vergeleken met een datum, een gegeven oplossing. Dit kan een al bestaande oplossing zijn. De verdere stappen zijn als volgt:

1. Vergelijk de eigenschappen van elk alternatief met die van het datum. De waardering is als volgt:
 - + = beter dan het datum
 - - = slechter dan het datum
 - s = even goed ('same') als het datum of 'een twijfelgeval'
2. Vergelijk en waardeer deze score-profielen. Veel plussen en en weinig minnen wijst op een goed alternatief.

Wanneer de stappen 1 en 2 toegepast worden op 4 criteria en 5 alternatieven, zou de volgende tabel kunnen ontstaan:

	A1	A2	A3	A4	A5	datum
C1	+	-	-	+	s	s
C2	-	-	+	s	+	s
C3	+	+	-	s	+	s
C4	s	s	s	-	-	s

Tabel 1: Voorbeeld van een Pugh beslissingsmatrix

In dit voorbeeld zijn A2 en A3 voor dit datum minder geschikt.

3. Kies een nieuw datum en herhaal de stappen 1 en 2.

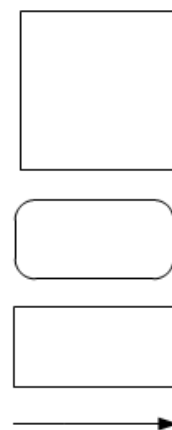
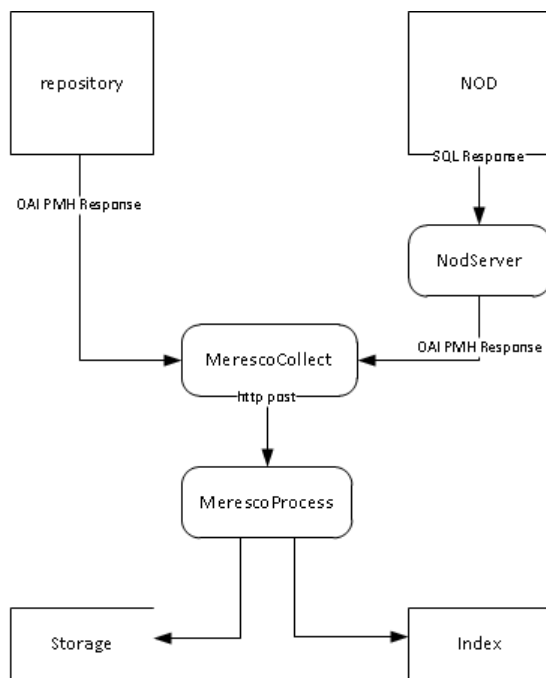
In bovenstaand voorbeeld zou A1 of A5 een leuk nieuw datum zijn.

4. Herhaal dit proces tot de beslissers overeenstemming bereiken over het beste alternatief.

Harvesten

Voor het harvesten van de metadata uit de verschillende repositories door NARCIS wordt gebruik gemaakt van Meresco, een open source product van SeekYouToo [OSS02]. Deze haalt via het OAI-PMH protocol¹ de metadata uit de repositories en bouwt een index. Ook wordt de metadata in een aantal formaten opgeslagen, waaronder het originele record. Dit verlaagt de hoeveelheid I/O bij het opvragen. Voor sommige toepassingen is niet het hele record nodig, maar bijvoorbeeld alleen de identifier, de titel en de auteur.

MerescoCollect bevraagt dagelijks de repositories en de NodServer. Bij de Request kan een ResumptionToken meegegeven worden om de nieuwste records te verkrijgen. Het ResumptionToken geeft aan waar het harvest-proces in de repository-set gebleven was. MerescoProcess ontvangt deze records via http post en verwerkt ze: ze slaat de velden die doorzocht moeten kunnen worden op in de index en slaat het record in verschillende formaten op in storage. Figuur 1 toont een schematische voorstelling van deze zaken, figuur 2 toont van boven naar beneden de volgende Gane Sarson Data Flow Diagram symbolen: externe input, proces, opslag, datastroom.



Figuur 2: de Gane Sarson symbolen

Figuur 1: Gane Sarson Data Flow Diagram

Repository

in de repositories staat de metadata van publicaties en datasets opgeslagen

NOD (Nederlandse Onderzoek Databank)

de data betreft onderzoekers, instituten en onderzoekbeschrijvingen en de onderlinge relaties.

Storage

filesysteem waarin elk record in de verschillende gewenste formaten wordt

¹ <http://www.openarchives.org/OAI/openarchivesprotocol.html>

opgeslagen. Elk record wordt in ieder geval het originele formaat opgeslagen en in know_short, een eigen NARCIS formaat.

Index

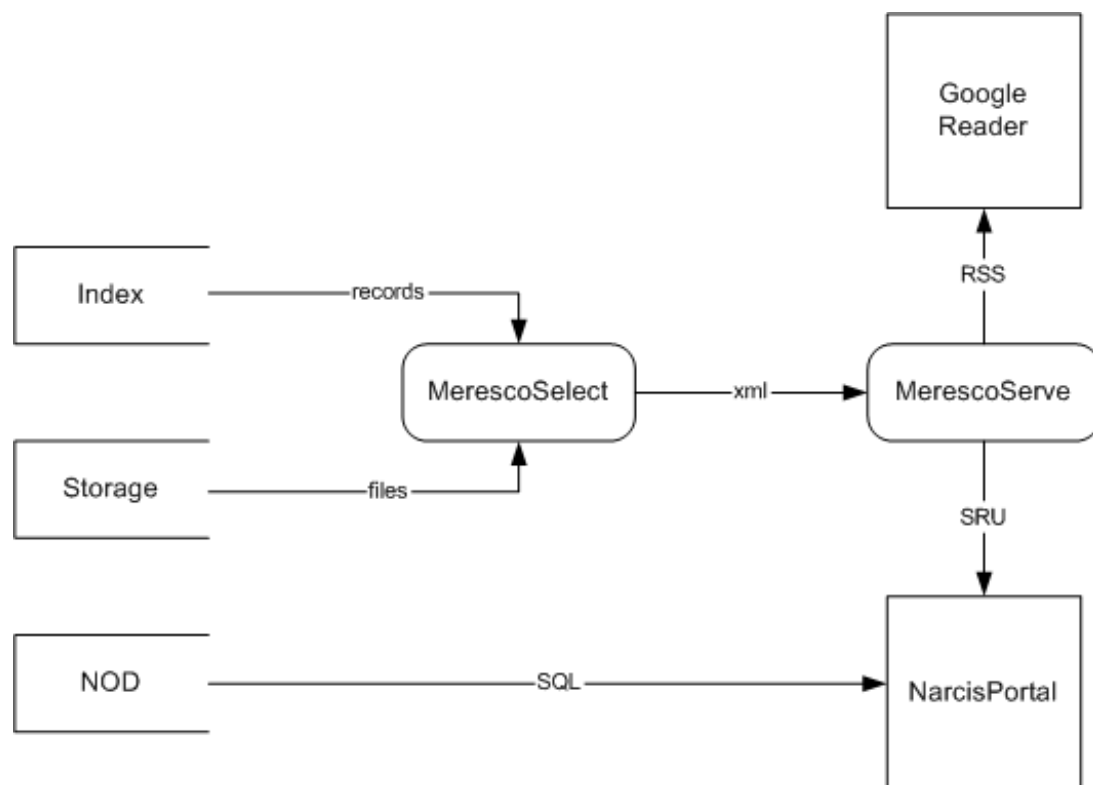
De index bevat alle data waarop gezocht kan worden. Net als bij *storage* is dit een versimpelde weergave van de werkelijkheid, aangezien er verschillende indexen zijn, bijvoorbeeld voor facetten, voor untokenized velden etc.

OAI-PMH Response

XML formaat van de Library of Congres om metadata uit te wisselen in antwoord op een OAI-PMH Request.

Http post

upload naar de index



Figuur 3: Het zoeken in de index en het uitleveren van de records

Bovenstaande figuur 3 toont welke datastromen een zoekvraag in de NARCIS Portal tot gevolg heeft. De SRU-vraag wordt beantwoord door MerescoSelect die in de Index de records opvraagt. De desbetreffende files worden uit Storage gehaald, tot een xml samengevoegd en in een SRU-response verpakt naar de NARCIS Portal gestuurd. Wanneer er data uit de NOD nodig is die niet in Storage staat, wordt dit rechtstreeks uit de NOD gehaald.

MerescoServe kan ook andere protocollen leveren, hier wordt een RSS-uitvoer getoond.

ATOM

Atom is als uitwisselingsprotocol een stuk armer dan OAI-PMH. Er kan niet aangegeven worden dat bepaalde records niet meer bestaan en ook het 'verder waar ik gebleven was' idee is hier niet verwerkt. De syndication feed bestaat uit een opsomming van entries, hopelijk chronologisch gesorteerd.

```
<?xml version="1.0" encoding="utf-8"?>
<feed xmlns="http://www.w3.org/2005/Atom">

  <title>ReMs at http://myrepository.org/</title>
  <link href="http://myrepository.org/" />
  <link href="http://myrepository.org/all-rems.atom" rel="self"/>
  <updated>2007-08-15T18:30:02Z</updated>
  <author>
    <name>John Doe</name>
    <email>johndoe@myrepository.org</email>
  </author>
  <id>urn:nbn:nl:ui:10-1234/5678</id>

  <entry>
    <title>Lorem Ipsum</title>
    <link rel="alternate" href="http://myrepository.org/resource/xyz.html"/>
    <link rel="resource" href="http://myrepository.org/resource/xyz.rdf"/>
    <id>urn:nbn:nl:ui:10-1234/xyz</id>
    <updated>2011-01-13T10:42:02Z</updated>
  </entry>

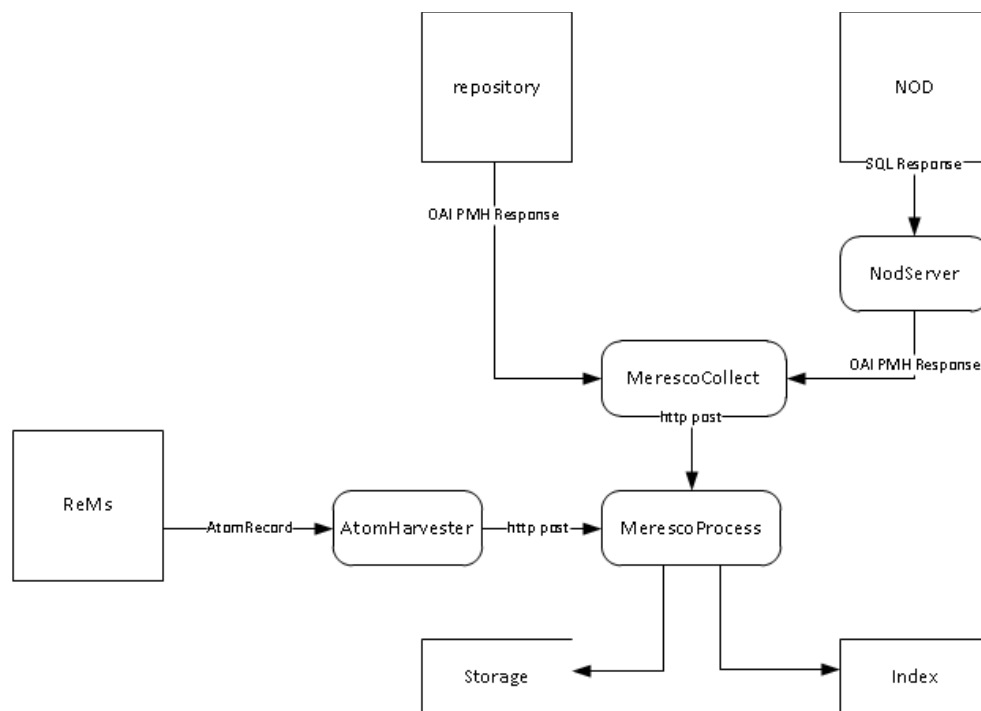
  <entry>
    <!-- ReM For Object2 -->
    ...
  </entry>

  <entry>
    <!-- ReM For Object3 -->
    ...
  </entry>

</feed>
```

*Figuur 1: all-rems.atom zoals beschreven op
<http://wiki.surffoundation.nl/display/vp/Discovery+and+Transport>*

Om in de bestaande flow ook de ATOM-records op te nemen, moet het proces zoals weergegeven in figuur 1 uitgebreid worden met een AtomHarvester. Dit is functioneel equivalent met MerescoCollect en voedt het MerescoProcess via de http post.



Figuur 2: AtomHarvester toegevoegd aan het Data Flow Diagram

ReMs

De Server die de Resource Maps (zoals getoond in figuur 1) serveert. Zie figuur 3 voor het atom protocol.

AtomHarvester

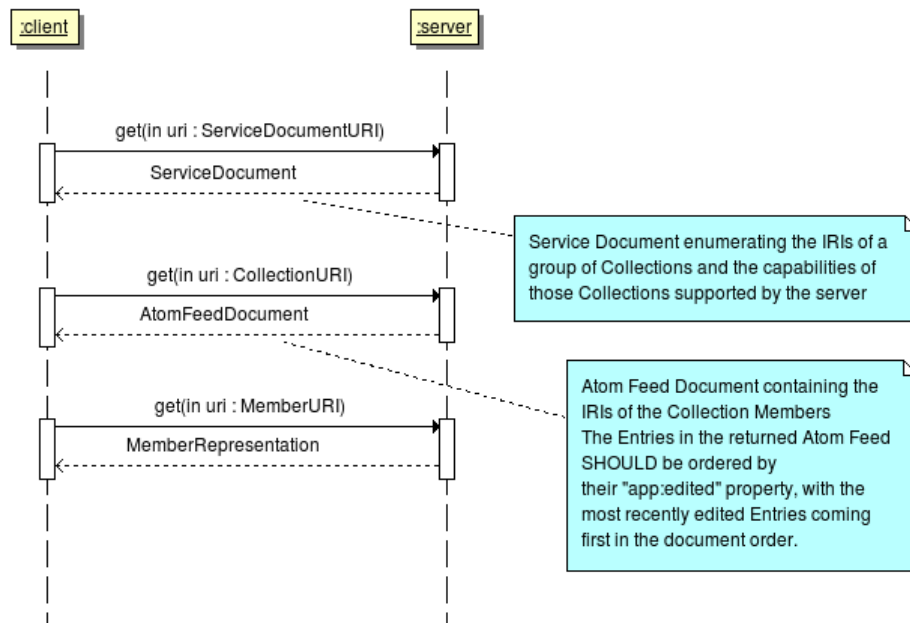
De AtomHarvester moet de administratieve taken die OAI-PMH regelt nu zelf oplossen. Het leest de (nieuwe) ReMs in, voegt metadata met betrekking tot de harvesttaak toe en post het naar het MerescoProcess

AtomRecord

Verrijkte Publicatie verpakt in een Atom Record

Http post

upload naar de index



Figuur 3: ATOM publishing request-response

Volledig RDF voorbeeld

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcterms="http://purl.org/dc/terms/" xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:dai="http://purl.org/info:eu-repo/dai/" xmlns:eurepo="http://purl.org/info:eu-
repo/semantics/"
  xmlns:wdrs="http://www.w3.org/2007/05/powder-s#"
  xmlns:ore="http://www.openarchives.org/ore/terms/"
  xmlns:escape-system="http://purl.utwente.nl/ns/escape-system.owl#"
  xmlns:escape-project="http://purl.utwente.nl/ns/escape-projects.owl#"
  xmlns:escape-agents="http://purl.utwente.nl/ns/escape-agents.owl#"
  xmlns:escape-pubtypes="http://purl.utwente.nl/ns/escape-pubtypes.owl#"
  xmlns:swan="http://swan.mindinformatics.org/ontologies/1.2/discourse-relationships/">
  <rdf:Description
    rdf:about="http://myrepo.org/rsrc/resourcemap/oai:dspace.library.uu.nl:1874/14036">
    <ore:describes
      rdf:resource="http://myrepo.org/rsrc/unapi?ReM=oai:dspace.library.uu.nl:1874/14036" />
    <dcterms:creator rdf:resource="http://www.narcis.nl/#remAgent" />
    <dcterms:modified>Mon Apr 11 15:07:35 CEST 2011</dcterms:modified>
  </rdf:Description>
  <rdf:Description
    rdf:about="http://myrepo.org/rsrc/unapi?ReM=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="info:eu-repo/dai/nl/16717231X" />
    <ore:aggregates rdf:resource="info:eu-repo/dai/nl/070442215" />
    <ore:aggregates rdf:resource="info:eu-repo/dai/nl/304830097" />
    <ore:aggregates rdf:resource="info:eu-repo/dai/nl/070442258" />
    <ore:aggregates rdf:resource="info:eu-repo/dai/nl/304831387" />
    <ore:aggregates rdf:resource="ORG1236845" />
    <ore:aggregates rdf:resource="OND1292696" />
    <ore:aggregates
      rdf:resource="oai:easy.dans.knaw.nl:twips.dans.knaw.nl--6454404375551860837-1192783132652" />
    <dcterms:title>Doctoral Thesis: The Development of Personality and
      Problem Behaviour in Adolescence</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 in
      Adolescence</dcterms:title>
    <escape-project:isOutcomeDocumentOf
      rdf:resource="OND1292696" />
    <foaf:makes rdf:resource="info:eu-repo/dai/nl/304831387" />
    <dcterms:description>This dissertation focuses on the development of
      personality and problem behaviour in adolescence. It has long been
      debated whether personality is stable or whether it changes over
      time. Personality could change due to the many changes that occur
      during adolescence, such as attending a new school, ge</dcterms:description>
  </escape-pubtypes:DoctoralThesis>
  <foaf:OnlineAccount rdf:about="URN1874">
    <rdf:type rdf:resource="http://www.persistent-identifier.nl" />
    <foaf:accountName>URN:NBN:NL:UI:10-1874-14036</foaf:accountName>
  </foaf:OnlineAccount>
  <escape-agents:AcademicStaff rdf:about="info:eu-repo/dai/nl/16717231X">
    <escape-project:worksOn rdf:resource="D54000" />
    <foaf:account>
      <rdf:type rdf:resource="http://purl.org/info:eu-repo/dai" />
      <foaf:accountName>info:eu-repo/dai/nl/16717231X</foaf:accountName>
    </foaf:OnlineAccount>
  </foaf:account>
```

```

<foaf:workplaceHomepage>http://www.fmg.uva.nl/communicatiewetenschap/</foaf:workplaceHomepage>
<foaf:workplaceHomepage>http://www.ru.nl/pwo</foaf:workplaceHomepage>
<foaf:workplaceHomepage>http://www.ru.nl/bsi/</foaf:workplaceHomepage>
<foaf:family_name>Engels</foaf:family_name>
<foaf:givenname>R.C.M.E.</foaf:givenname>
<dcterms:title>R.C.M.E. Engels</dcterms:title>
<foaf:pastProject rdf:resource="OND1292696" />
</escape-agents:AcademicStaff>
<escape-project:Topic rdf:about="D54000">
  <escape-project:isWorkedOnBy rdf:resource="info:eu-repo/dai/nl/304830097" />
  <escape-project:isWorkedOnBy rdf:resource="info:eu-repo/dai/nl/16717231X" />
  <escape-project:isWorkedOnBy rdf:resource="info:eu-repo/dai/nl/070442258" />
  <escape-project:isWorkedOnBy rdf:resource="info:eu-repo/dai/nl/070442215" />
  <dcterms:title>Pedagogics</dcterms:title>
</escape-project:Topic>
<escape-agents:AcademicStaff rdf:about="info:eu-repo/dai/nl/070442215">
  <escape-project:worksOn rdf:resource="D61000" />
  <escape-project:worksOn rdf:resource="D54000" />
  <escape-project:worksOn rdf:resource="A84100" />
  <escape-project:worksOn rdf:resource="A84200" />
  <foaf:family_name>Meeus</foaf:family_name>
  <foaf:family_name>W. Meeus</foaf:family_name>
  <foaf:givenname>W.H.J.</foaf:givenname>
  <dcterms:title>W.H.J. Meeus</dcterms:title>
  <dcterms:title>W. Meeus</dcterms:title>
  <foaf:pastProject rdf:resource="OND1292696" />
  <foaf:member rdf:resource="ORG1237506" />
</escape-agents:AcademicStaff>
<escape-project:Topic rdf:about="D61000">
  <escape-project:isWorkedOnBy rdf:resource="info:eu-repo/dai/nl/070442215" />
  <dcterms:title>Sociology</dcterms:title>
</escape-project:Topic>
<escape-project:Topic rdf:about="A84100">
  <escape-project:isWorkedOnBy rdf:resource="info:eu-repo/dai/nl/070442215" />
  <dcterms:title>Education</dcterms:title>
</escape-project:Topic>
<escape-project:Topic rdf:about="A84200">
  <escape-project:isWorkedOnBy rdf:resource="ORG1236845" />
  <escape-project:isWorkedOnBy rdf:resource="info:eu-repo/dai/nl/070442215" />
  <dcterms:title>Education</dcterms:title>
</escape-project:Topic>
<escape-agents:AcademicStaff rdf:about="info:eu-repo/dai/nl/304830097">
  <escape-project:worksOn rdf:resource="D54000" />
  <foaf:family_name>Hale</foaf:family_name>
  <foaf:givenname>W.W.</foaf:givenname>
  <dcterms:title>W.W. Hale</dcterms:title>
  <foaf:pastProject rdf:resource="OND1292696" />
  <foaf:member rdf:resource="ORG1237506" />
</escape-agents:AcademicStaff>
<escape-agents:AcademicStaff rdf:about="info:eu-repo/dai/nl/070442258">
  <escape-project:worksOn rdf:resource="A82000" />
  <escape-project:worksOn rdf:resource="A84000" />
  <escape-project:worksOn rdf:resource="D54000" />
  <foaf:family_name>Raaijmakers</foaf:family_name>
  <foaf:givenname>Q.A.W.</foaf:givenname>
  <dcterms:title>Q.A.W. Raaijmakers</dcterms:title>
  <foaf:pastProject rdf:resource="OND1292696" />
  <foaf:member rdf:resource="ORG1237506" />
</escape-agents:AcademicStaff>
<escape-project:Topic rdf:about="A82000">
  <escape-project:isWorkedOnBy rdf:resource="info:eu-repo/dai/nl/070442258" />
  <dcterms:title>Social circumstances</dcterms:title>
</escape-project:Topic>

```

```

<escape-project:Topic rdf:about="A84000">
  <escape-project:isWorkedOnBy rdf:resource="info:eu-repo/dai/nl/070442258" />
  <dcterms:title>Education and training</dcterms:title>
</escape-project:Topic>
<escape-agents:PhDCandidate rdf:about="info:eu-repo/dai/nl/304831387">
  <foaf:family_name>Akse</foaf:family_name>
  <dcterms:title>J. Akse</dcterms:title>
  <foaf:givenname>J.</foaf:givenname>
  <foaf:pastProject rdf:resource="OND1292696" />
</escape-agents:PhDCandidate>
<escape-agents:University rdf:about="ORG1236845">
  <escape-project:worksOn rdf:resource="A90000" />
  <escape-project:worksOn rdf:resource="A84200" />
  <foaf:name>Universiteit Leiden</foaf:name>
</escape-agents:University>
<escape-project:Topic rdf:about="A90000">
  <escape-project:isWorkedOnBy rdf:resource="ORG1236845" />
  <dcterms:title>Fundamental research</dcterms:title>
</escape-project:Topic>
<foaf:Project rdf:about="OND1292696">
  <escape-system:resourceUri
    rdf:resource="%3Fwicket%3AbookmarkablePage%3D
%3Anarcis.portal.class+nl.know.oi.nod.records.ResearchPage%26RecordID%3DOND1292696" />
  <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:maker rdf:resource="ORG1236845" />
  <foaf:maker rdf:resource="ORG1237506" />
</foaf:Project>
<foaf:Document
  rdf:about="oai:easy.dans.know.nl:twips.dans.know.nl--6454404375551860837-1192783132652">
  <swan:relatedTo rdf:resource="oai:dspace.library.uu.nl:1874/14036" />
  <dcterms:title>CONAMORE (Conflict and Management of Relationships)
    Adolescent study</dcterms:title>
  <foaf:maker rdf:resource="info:eu-repo/dai/nl/070442215" />
  <dcterms:description>CONAMORE is an ongoing longitudinal study of
    Dutch adolescents that examines
    their personality and identity, their relationships with parents and peers
    as well as
    their emotional and behavioural states (Meeus et al., 2002). Between 2001
    and
    2006 five annual waves have been</dcterms:description>
</foaf:Document>
</rdf:RDF>

```

Afkortingenlijst

De meerwaarde van de Digital Author Identification

BIK

Beleidsgroep Innovatie & Kennisinfrastructuur is een door SURF geïnitieerde adviesgroep voor het landelijke platformbestuur ICT en Onderzoek. Hierin zijn vanuit de universiteiten de bibliothecarissen, informatiemanagers en beleidsmedewerkers vertegenwoordigd. Een greep uit de recente agendapunten: CRIS, *DAI*, DOI

CRIS

Current Research Information System

CQL

Common Query Language

DAI

Digital Author Identification. Een uniek number, toegekend door *OCLC-PICA* aan elke Nederlandse onderzoeker. OCLC heeft in de *NTA* aan Nederlandse auteurs een *PPN* toegekend. Wanneer voor een onderzoeker een aanvraag voor een DAI binnenkomt, wordt één van de PPN's tot DAI uitverkoren. In de praktijk blijkt dat er niet altijd een 1-op-1 relatie bestaat tussen DAI en onderzoeker. Er zijn onderzoekers met meer dan één DAI, en er zijn ook DAI-nummers met meer dan één onderzoeker. Er wordt gewerkt om dit te 'ontdubbelen', maar dat is een tijdrovend proces. Niet alleen het opsporen, maar ook om alle afgeleide bestanden weer te synchroniseren met de bron.

Dit is een erfenis van de NTA waar deze problemen met de PPN's ook al bestonden.

DANS

Data Archiving and Networked Services, een instituut van de KNAW en NWO.

demonstrator

niet te verwarren met een mockup. Een demonstrator is een werkend product, waarin de functionaliteit zichtbaar is.

In deze context equivalent met prototype

DIDL

Digital Item Declaration Language

XML schema voor de metadata van digitale records, zoals publicaties. Gebruik in Nederland grotendeels in combinatie met MODS

DOI

Digital Object Identifier, internationale implementatie van de Persistent Identifier. Nederland gebruikt hiervoor een URN.

DRIVER

Digital Repository Infrastructure Vision for European Research, NARCIS op Europese schaal. De NARCIS Repository zal als nationale repository voor DRIVER fungeren.

Dublin Core

formaat voor metadata bestanden. verplicht beschikbaar bij het OAI-PMH protocol.
Gebruik in Nederland grotendeels vervangen door DIDL/MODS

GGC

Gemeenschappelijke Geautomatiseerde Catalogus. Een product van OCLC PICA waar bibliotheken gegevens in opslaan over auteurs, boeken etc.

De NTA is een onderdeel van de GGC? [QQQ]

harvesten

het 'oogsten' van de repositories. Dit gaat aan de hand van het OAI-PMH protocol.

OAI-PMH schrijft voor dat elk item tenminste in Dublin Core geretrieved kan worden.

KNAW

Koninklijke Nederlandse Academie van Wetenschappen

keurling

Selecte groep waarvan alle publicaties (dus niet alleen open access publicaties) werden opgenomen in de repositories. Ingevoerd om DAREnet te promoten.

Oude publicaties zijn in dit verband soms ingescand

MARC

MAchine Readable Cataloging.

Catalog systeem ontwikkeld door de Library of Congress in de jaren zestig om bibliografische data te beschrijven.

Metis

Metis is de Nederlandse implementatie van een *CRIS*. Alle 13 Nederlandse universiteiten hebben een Metis, waarin metadata mbt de onderzoekers, projecten en publicaties bewaard wordt.

MODS

Metadata Object Description Schema

XML schema ontwikkeld door de Library of Congress. Ligt tussen het complexe *MARC* en het overgesimplificeerde *Dublin Core*.

NARCIS Index

het aggregaat van de repositories van de universiteiten, een aantal onderzoeksinstituten en de NOD. Handelt zoekvragen af via SRU/CQL

NARCIS Portal

website die de relaties tussen onderzoek, onderzoekers en publicaties inzichtelijk maakt. Showcase voor de NARCIS Index.

NARCIS Repository

storage die door derden, ook internationaal, gebruikt kan worden als Nationale Repository

NARCIS Suite

samenvoeging van de NARCIS Index, Portal en Repository.

NOD

Nederlandse Onderzoek Databank. Database met onderzoekinformatie betreffende wetenschappelijke instituten, onderzoekers en onderzoeken, beheerd door de afdeling OnderzoekInformatie.

NTA

Nederlandse Thesaurus voor Auteursnamen
onderhouden door de samenwerkende bibliotheken, georganiseerd in *OCLC PICA*, met het doel catalogi samen te stellen. Elke auteur krijgt hier een voorkeursnaam en een pica productienummer (*PPN*), wat later als invoer is gebruikt voor de *DAI*.

NWO

De Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) financiert duizenden toponderzoekers aan universiteiten en instituten en geeft sturing aan de Nederlandse wetenschap via subsidies en onderzoeksprogramma's.

OAI-PMH

Open Archives Initiative - Protocol for Metadata Harvesting

Een http based protocol waarmee repositories *geharvest* kunnen worden. In ieder geval moet *Dublin Core* ondersteund worden. Een aantal verbs die gebruikt kunnen worden: GetRecord, ListIdentifiers, ListRecords.

zie <http://www.openarchives.org/pmh/>

OCLC-PICA

OCLC PICA is een internationaal samenwerkingsverband tussen wetenschappelijke bibliotheken.

Open Access Publicaties

Publicaties waarop geen embargo (meer) rust. Vrij toegankelijk.

Persistent Identifier

Identificatie die altijd beschikbaar is en via een resolver verwijst naar het gerefereerde object. Internationaal gerealiseerd door de Digital Object Identifier, nationaal door een URN.

PPN

Pica Productienummer. Door *OCLC PICA* uitgedeelde nummers aan wetenschappelijke auteurs. Hierbij is niet heel zorgvuldig te werk gegaan, sommige nummers zijn vaker dan één keer uitgedeeld. Op het moment dat een OCLC-bibliotheek een PPN aanvraagt, wordt die toegekend. De meeste auteurs hebben meer dan één nummer, bijvoorbeeld omdat een persoon vaker is ingevoerd en pas later tot één record samengevoegd.

Een onderdeel van de *DAI* uitrol was het promoveren van één PPN per auteur tot *DAI*. Helaas zijn soms meerdere PPNs per auteur gepromoveerd, waardoor de data bevuild is.

RDF

triples (Subject - Predicaat - Object) die tezamen een object beschrijven. Bij *Enhanced Publication* gebruikt in combinatie met FOAF, ORE, DCTERMS om de predicaten te definiëren.

repository

opslag van publicatie en bijbehorende metadata. Ook gebruikt voor retrieval.

resolver

key-value mapping. In dit verband: mapping van een Persistent Identifier naar het bijbehorende object.

SPARQL

Sparql Protocol And RDF Query Language. Query language

SRU

het Search/Retrieve via URL protocol

Verrijkte Publicatie

Enhanced Publication. Een RDF XML representatie van een publicatie (of ander object) met bijbehorende, samenstellende onderdelen.

WISH

Wijzigingen in de metadatastructuur worden voorbereid binnen de zg. WISH-groep en vervolgens besproken binnen het *WRM*-overleg. Daarin worden besluiten genomen ten aanzien van invoering hiervan.

WRM

Wergroep Repository Managers maakt afspraken over formaten, functionaliteit etc met betrekking tot de repositories van de universiteiten

Plan van Aanpak

de meerwaarde van de Digital Author Identification

Inhoudsopgave

1Projectdefinitie.....	2
1.1Situatie.....	2
1.2Vraagstelling.....	3
1.3Scope.....	3
1.4Projectrisico's.....	3
2Projectaanpak.....	4
2.1Activiteiten.....	4
2.2Projectorganisatie.....	5
2.3Producten en planning.....	6
3Projectbeheersing.....	7
3.1Kwaliteit.....	7
3.2Beheersing van de projectrisico's.....	7

1 Projectdefinitie

1.1 Situatie

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."* [http://www.knaw.nl ; 2011-02-05]

Dit project vindt plaats binnen de afdeling Onderzoek Informatie (OI). Deze afdeling bestaat uit 8FTE en houdt zich bezig met het toegankelijk en transparant maken van Nederlands, publiek-gefinancierd onderzoek en het zichtbaar maken van kennis van onderzoekers en onderzoeksinstituten. Dit gebeurt onder andere met het product NARCIS.

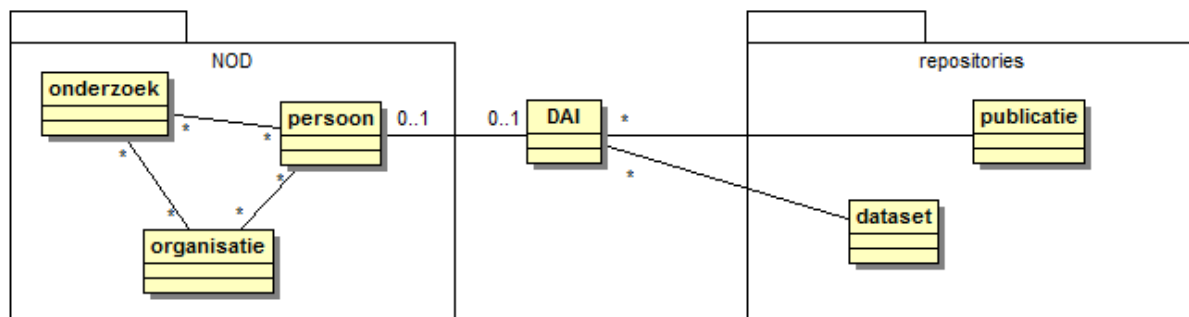
NARCIS haalt de data uit de repositories[*]¹ van alle Nederlandse universiteiten en een groot aantal onderzoeksinstituten. Bovendien heeft NARCIS toegang tot de Nederlandse Onderzoek Databank (NOD), waar alle onderzoek waar NWO-geld naar toe gaat, is aangemeld.

SURFfoundation heeft een aantal jaar geleden de DAI ingevoerd, de Digital Author Identification. Dit is een uniek landelijk nummer dat elke auteur, die verbonden is aan een Nederlandse universiteit of onderzoeksinstituut kan aanvragen.

Langzaam maar zeker vindt de DAI zijn weg in de repositories. Hierdoor is het momenteel mogelijk alle publicaties van eenzelfde auteur te vinden, ongeacht of de naam van de auteur anders geschreven is, of dat de publicaties in andere repositories staan.

Deze publicatielijsten zijn nu voor iedereen toegankelijk via de Portal : www.narcis.nl.

Figuur Error: Reference source not found toont het datamodel waar de Portal uit put.



Figuur 1: versimpelde weergave van de DAI als verbinding tussen de repositories en de NOD

De afdeling OI vermoedt dat de koppeling van de data tussen de repositories voor nog veel meer toepassingen gebruikt zou kunnen worden. Al helemaal wanneer de data uit de NOD meegenomen wordt.

Op deze manier kan NARCIS zich profileren als nationaal focal point voor wetenschappelijke publicaties en onderzoeksinformatie.

¹ Zie bijlage 2

De positie van de Nederlandse kennisinfrastructuur is uniek in Europa. Nergens anders is de samenwerking tussen alle universiteiten zo nauw en worden afspraken over de formaten en technieken door allen nagestreefd. NARCIS vervult hierin een rol als nationale aggregator. Om deze positie te behouden moet ze zorgen dat de data beschikbaar is en op een toegankelijke manier gepresenteerd wordt.

De Europese collega's hebben ook belangstelling voor de DAI. De afdeling wil graag laten zien wat met zo'n koppeling allemaal mogelijk is.

NARCIS wordt een krachtiger tool en geeft een completer beeld wanneer de DAI bij meer publicaties wordt aangegeven. Om dit te bereiken, moeten de repository managers deze data ter beschikking willen stellen.

1.2 Vraagstelling

De afdeling Onderzoek Informatie wil weten op welke manier de meerwaarde van de DAI het beste aangetoond kan worden aan de repository managers. Ze wil een demonstrator ontwikkelen om draagvlak te creëren. Ook wil ze aan de Europese collega's laten zien wat de voordelen zijn van de Nederlandse kennisinfrastructuur en de DAI.

1.3 Scope

De afdeling Onderzoek Informatie kent de wensen van de eindgebruikers door het houden van gebruikersonderzoek. Dit is recentelijk uitgevoerd. Nu is ze geïnteresseerd in de eisen en wensen van de leveranciers van de data. De repository manager zal dus een voorname rol spelen.

Het ontwikkelen van een demonstrator is onderdeel van het project. De beslissing over al of niet in productie nemen van de opgeleverde code is geen onderdeel van het project. Deze beslissing ligt bij het NARCIS management.

1.4 Projectrisico's

1. binnenkort gaat de afdeling Onderzoek Informatie over van het Bureau van de KNAW naar DANS, een ander instituut van de KNAW. Dit betekent, naast een andere werkplek, infrastructuur, collega's, ook een andere opdrachtgever, en mogelijk een andere bedrijfsmentor.
2. de afdeling blijft tijdens het project doorwerken aan nieuwe ontwikkelingen. Dit zou kunnen interfereren met dit project, bijvoorbeeld als men overgaat op een nieuwe versie van de backend.
3. Voor het onderzoek moeten externen gesproken worden. Dit moet al op korte termijn gebeuren. Wanneer dit niet tijdig kan plaatsvinden, duurt het onderzoek langer dan gepland, of is de uitkomst niet optimaal.
4. Dit project heeft een harde deadline. Dat betekent dat de tijd bij alle beslissingen een belangrijke factor is. Wanneer dit in het begin te weinig meespeelt, komen latere fasen in tijdsnood en worden de activiteiten met minder kwaliteit uitgevoerd.
5. De functionaliteit van de te bouwen demonstrator moet nog bepaald worden. Wanneer de eisen niet duidelijk zijn, bestaat het risico dat de uiteindelijke demonstrator niet aan de nog-niet-omschreven verwachtingen voldoet.

2 Projectaanpak

Het project is onderverdeeld in twee gescheiden deelprojecten, die aansluitend uitgevoerd worden.

2.1 Activiteiten

2.1.1 deelproject 1: Onderzoek

In de inception fase vindt een onderzoek plaats naar de mogelijkheden van de DAI. Dit bestaat uit een aantal deelactiviteiten:

- aanscherpen onderzoeksvraag
- divergeren:
 - identificeren van de stakeholders
 - inventariseren en wegen van de wensen en eisen van de stakeholders dmv gesprekken, verslagen van vergaderingen
 - bedenken van oplossingsmogelijkheden, binnen de grenzen van de afdeling, technische beperkingen, tijdslimiet van het project
- convergeren
 - oplossingsmogelijkheden scoren aan de hand van de wensen van de stakeholders
 - weging van de eisen van de stakeholders toepassen op de oplossingsmogelijkheden
 - in overleg met de opdrachtgever maken van een ranking van de oplossingsmogelijkheden

2.1.2 deelproject 2: Ontwikkelen

Het tweede deel van het project bestaat uit het ontwikkelen van de demonstrator.

Als ontwikkelmethode heb ik gekozen voor een iteratieve methode, RUP.

De demonstrator wordt ontwikkeld in een Elaboration en Construction fase.

- Elaboration
 1. software architecture document
 2. design model
 3. data model
 4. testing mechanisms (worst-case scenario's)
 5. prototype
- Construction
 1. the system, volgens TDD
 2. testing results

Ten behoeve van de overdracht wordt de demonstrator gepresenteerd, bijvoorbeeld in de Werkgroep Repository Managers (WRM) of aan de afdeling Onderzoek Informatie. De activiteiten die hierbij uitgevoerd moeten worden:

1. voorbereiden presentatie
2. geven presentatie / demonstratie
3. inventariseren feedback

2.2 Projectorganisatie

Bij de opzet van het afstudeerplan is Marga van Meel betrokken geweest, het afdelingshoofd van Onderzoek Informatie. Per 8 februari 2011 gaat de afdeling op in DANS, met Henk Harmsen als Adjunct Directeur.

Op een aantal momenten zal er overleg zijn tussen opdrachtgever en opdrachtnemer:

- Er is een wekelijks overleg met de bedrijfsmentor en een driewekelijks overleg met de opdrachtgever. Op deze manier blijven zij betrokken en zijn er geen verrassingen.
- Het plan van aanpak wordt besproken met en goedgekeurd door de opdrachtgever.
- Het onderzoeksresultaat wordt besproken met en goedgekeurd door de opdrachtgever.

2.3 Producten en planning

De activiteiten resulteren met iteraties in de volgende producten:

	activiteit	product	planning	
Inception fase				
1	onderzoek	Plan van Aanpak	4	
2		onderzoekrapport	15	
3		shortlist		
4		globale planning	2	
Elaboration fase				
5	in kaart brengen huidige situatie	software architecture doc	5	3
6		design model	3	
7		data model	2	
8		prototype	2	2
9		worst-case scenario's	2	
Construction fase				
10	TDD	systeem	10	7
11	testen	testresultaten	2	2
overdracht				
12	voorbereiden presentatie	presentatie	3	
13	inventariseren feedback	feedback	1	
14		conclusies & aanbevelingen	2	
portfolio				
15		afstudeerverslag	15	

*Tabel 1: producten met de geplande duur in dagen,
de elaboration en construction fase kennen twee iteraties*

3 Projectbeheersing

3.1 Kwaliteit

- Het onderzoek zorgt voor een goede onderbouwing van de te realiseren demonstrator. Het garandeert dat de stakeholders erbij betrokken worden en creëert draagvlak voor eindresultaat
- Door tijdens de Elaboration en Construction fase te werken volgens Test Driven Development wordt gestuurd op het voldoen aan de requirements. TDD zorgt ervoor dat er altijd een werkend stuk code ligt, dat in ieder geval deels de gevraagde functionaliteit levert.
- Door het uitvoeren van unittesten, integratie-testen en functionele acceptatietesten wordt de kwaliteit van het product vastgesteld.
- De tijd die beschikbaar is voor het project is beperkt tot de duur van het afstuderen. Om de kwaliteit van de uitgevoerde werkzaamheden te waarborgen, zal dit in het onderzoek meegenomen worden als een van de eisen voor de demonstrator.
- Door in de construction fase volgens TDD en incrementeel/iteratief te werken, is er altijd een werkende applicatie. Bovendien is er eventueel de mogelijkheid om te besluiten slechts 1 demonstrator te ontwikkelen.

3.2 Beheersing van de projectrisico's

1. [overstap naar DANS]
in een vroeg stadium de overdracht tussen beide opdrachtgevers regelen.
2. [branching]
Wanneer de wijzigingen beperkt zijn, is het mogelijk de wijzigingen te verwerken in de projectcode. Eventueel kan er gekozen worden om te werken met een bevroren versie.
3. [externen]
 - introductie vanuit KNAW
 - externe vergaderingen bijwonen
4. [tijdslimiet]
 - RUP (iteratief werken)
 - gebruik van proven methode TDD
 - stevige deadlines
 - afspraken na komen
5. [juiste weg]
 - heldere communicatie
 - afstemmen met opdrachtgever en school
 - tussentijdse evaluatie
6. [draagvlak]

- betrekken van repository managers
- wanneer mogelijk een presentatie op de WRM
- communicatie op de afdeling over de voortgang
- collega's betrekken

Gespreksverslagen stakeholders

verslag gesprek drs. Elly Dijk

voormalig voorzitter van de WRM

bij NARCIS betrokken als contactpersoon met de RMs

lid van de Narcis Advies Raad

maandag 14 februari 2011, 15:00

onderwerp: oriënterend gesprek over de visie van de beleidsmedewerkers op de aggregatie functie van NARCIS in het algemeen en de DAI in het bijzonder.

universiteiten zijn geïnteresseerd in juiste en complete data mbt hun onderzoekers. De repository managers proberen een consistente repository te krijgen/behouden. Bovendien zijn zij bezig met de gebruikscijfers en de zichtbaarheid van de content. Dit is niet de focus van de beleidsmakers. Zij zien NARCIS als een etalage voor hun onderzoekers. Zij zouden deze etalage-functie graag verder uitgebouwd zien. In een oogopslag een indruk krijgen van de wetenschappelijke activiteiten van een persoon. Of door een zoekfunctie naar personen op een specifiek onderwerp. Of door NARCIS te koppelen met andere bronnen, ook internationaal. En met publicaties waar vanwege de uitgever nog embargo op rust, bijvoorbeeld Elsevier.

Bepaalde onderzoeksgebieden, zoals geneeskunde, hebben een eigen specifieke repository, die los staat van de universiteiten. Ook data die hier staat zou idealiter op de NARCIS pages getoond worden.

Een deel van de universiteiten ziet hun repository voornamelijk als bewaarplaats voor de publicaties en metadata. Zij zijn bereid data ter beschikking te stellen in DC via OAI-PMH, maar besteden hier verder geen aandacht aan. Zij zijn geen lid van de WISH of WRM.

Een deel is veel ambitieuzer en is actief bezig met formaten, Europese projecten, cris-NL etc. Zij hebben ook een betere connectie met hun Metis. Hun verwachtingen van NARCIS zijn navenant hoger.

verslag gesprek met drs. A. Th. Hogenaar

voormalig repository beheerder van de KNAW

lid van de WRM

maandag 21 februari 2011, 15:00

onderwerp: oriënterend gesprek over de visie van de repository managers op de aggregatie functie van NARCIS in het algemeen en de DAI in het bijzonder.

repository managers (rm) zijn geïnteresseerd in complete overzichten van hun onderzoekers. In de eigen repository zitten alleen de publicaties die bij dat instituut zijn uitgebracht. Wanneer een publicatie in samenwerking met een ander instituut is geschreven, komt het soms bij dat andere instituut in de repository.

Als onderzoekers bij een ander instituut gewerkt hebben, of daar een deel van hun werk uitvoeren, staan de publicaties dus verspreid over verschillende repositories.

De rms zien dit probleem, en zien de aggregerende taak van NARCIS vooral in dit licht. Ook het aanvullen van de data die bij een persoon gepubliceerd wordt met gegevens uit andere databanken zoals PubMed of LinkedIn ziet hij als een meerwaarde die NARCIS zou kunnen bieden.

Dhr. Hogenaar merkt op dat in de plannen van SURF de mogelijkheid bestond voor individuele onderzoekers om buiten de repository om publicaties in NARCIS in te voeren. Hij ziet wel het voordeel van de extra (en completere) data, maar alleen als de publicatie niet in de eigen (universitaire) repository thuis hoort. Bij de WISH is afgesproken dat alleen wetenschappelijke publicaties in de repo's geplaatst mogen worden. Niet elke universiteit houdt zich daaraan, waardoor bij sommige onderzoekers ook krantenartikelen en ander 'licht' materiaal genoemd worden. Deze restrictie zou vervallen bij zo'n by-pass.

Daarnaast willen de rms graag een overzicht van het verleden van hun onderzoekers, dus waar ze werkzaam geweest zijn.

Dhr. Hogenaar wijst erop dat deze data deels in de GGC aanwezig is, zij het onvolledig.

Hierin staan auteurs, met eventuele PPN's. Recentelijk is er een webinterface gemaakt, waarop adhv naam, titel etc gezocht kan worden in deze catalogus. De uitvoer kan onder andere in DAI XML geëxporteerd worden.

Dhr. Hogenaar waarschuwt dat OCLC geen goede naam heeft op het gebied van standaarden.

Onderzoeksrapport

De meerwaarde van de Digital Author Identification

opdrachtgever	Henk Harmsen
bedrijfsmentor	Chris Baars
opdrachtnemer	Linda Reijnhoudt
expert examiner	Gert Jan Timmerman
begeleider examiner	Tim Cocx

datum	02-05-11
versie	1.1

Inleiding

Dit rapport is het resultaat van het onderzoek naar de mogelijke toepassingen van de DAI in het kader van het afstudeerproject *De meerwaarde van de Digital Author Identification* aan de Haagse Hogeschool.

Hoofdstuk 1 poneert de vraagstelling. Hoofdstuk 2 behandelt de achtergrond, de processen, de data. Kortom, het hoe en wat.

Hoofdstuk 3 identificeert de verschillende stakeholders. Voor wie wordt deze demonstrator gebouwd? Hoofdstuk 3.5 beschrijft waarin de stakeholders geïnteresseerd zijn. Wat zijn hun eisen en wensen. In hoofdstuk 4 worden de mogelijke toepassingen van de DAI beschreven. Hoofdstuk 5 legt de toepassingen naast de eisen en wensen van de stakeholders. Het resultaat hiervan is een sortering van de toepassingen op wenselijkheid.

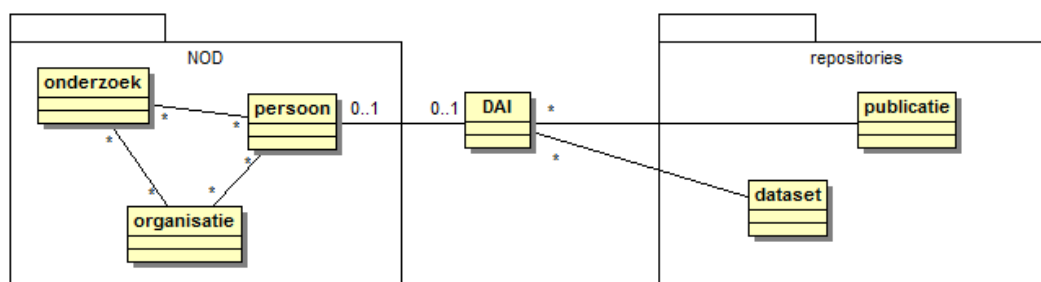
Ten slotte worden in hoofdstuk 6 de bevindingen gebruikt om een antwoord te formuleren op de hoofdvraag. Dit mondt uit in een advies.

Inhoudsopgave

1Vraagstelling.....	4
2Achtergrond.....	6
2.1De dataflow	6
2.2Verrijkte Publicaties.....	8
2.3De invoering van de DAI	9
3Stakeholders en hun wensen.....	10
3.1Repository Managers	10
3.2Management.....	10
3.3NARCIS projectleiding.....	10
3.4Uitvoerder.....	10
3.5wensenlijst.....	11
3.6De wensen gewogen door de stakeholders.....	12
4Potentiële toepassingen.....	13
4.1Mogelijkheden en beperkingen.....	13
4.2Lijst van potentiële toepassingen.....	13
4.3Potentiële toepassingen volgens de wensen.....	17
5De stakeholders en de potentiële toepassingen.....	18
6Conclusie.....	19
7Literatuur.....	19
8Bijlagen.....	20
8.1Cijfers huidige stand van zaken m.b.t. de DAI.....	20
8.2Verklarende woordenlijst.....	21
8.3DC	21
8.4DIDL / MODS XSD.....	21
8.5NOD Class Diagram.....	22
8.6PICA's DAI XML.....	24
8.7berekening weegfactoren criteria.....	25
8.8gespreksverslagen.....	25
8.9Privacy Statement GGC.....	25
8.10Kick-off Verrijkte Publicaties.....	25
8.11Precision en Recall.....	26
8.12NRC: Dit doen we goed.....	27

1 Vraagstelling

De afdeling Onderzoek Informatie heeft zich hard gemaakt voor de invoering van de Digital Author Identification (DAI). De DAI is de verbindende factor tussen de Nederlandse Onderzoek Databank (NOD) enerzijds en de repositories anderzijds. Publicaties en datasets van repositories onderling kunnen door de DAI verbonden worden. Illustratie 1 toont een versimpelde weergave van de classes binnen de NARCIS Portal waarin deze verbinding centraal staat.



Figuur 1: versimpelde weergave van de DAI als verbinding tussen de classes binnen de NARCIS Portal

De NARCIS Portal toont de data uit de repositories en de NOD. Op een persoonspagina bijvoorbeeld wordt data uit de NOD getoond, zoals de werkbetrekkingen van de persoon en contact informatie. Daarnaast wordt ook een publicatielijst getoond aan de hand van de DAI. Vaak is deze lijst nog incompleet. Niet alle publicaties zitten in de repositories en van lang niet alle publicaties zijn de auteurs voorzien van een DAI.

PERSON

PROF.DR. H.J. BENNIS



Main

Current research (4)

Completed research (10)

Publications (144)

Update Persondata >

Expertise

Expertise (NL)

Digital Author ID

Addition

Microvariation; (Generative) Syntax; Morphosyntax; Syntax-Semantics Interface; Dialectology

Microvariatie; (Generatieve) Syntaxis; Morphosyntaxis; Syntaxis-Semantiek Interface; Dialectologie

info:eu-repo/dai/nl/071792279

Bijzonder hoogleraar vanwege de Stichting Meertens; Directeur van het Meertens Instituut, Amsterdam

ACTIVE AS

Extraordinary professor

Organisation

Chair (EN)

Chair (NL)

Phone

Email

> Dutch Linguistics (UvA)

Dutch language variation

Taalvariatie binnen het Nederlands

+31-20-4628523

Hans. Bennis -at- meertens. knaw. nl

Management

Organisation

Phone

Email

URL

> Meertens Institute Research and documentation of Dutch language and culture (KNAW)

+31-20-4628523

Hans. Bennis -at- meertens. knaw. nl

> http://www.meertens.knaw.nl/cms/index.php?option=com_content...

Figuur 1: Persoonspagina van professor dr. H.J. Bennis

De repositories bestaan al geruime tijd, de DAI is pas een paar jaar geleden geïmplementeerd. Het achteraf invoeren van deze data in de repositories is voor sommige repositories een tijdrovende bezigheid.

Om het management van wetenschappelijke instellingen en universiteiten te overtuigen van de toegevoegde waarde van de DAI zal deze meerwaarde inzichtelijk gemaakt moeten worden. Dan kunnen zij beter geïnformeerd besluiten of hier resources voor beschikbaar gesteld zullen worden.

De hoofdvraag van dit onderzoek luidt dan ook:

[uit pva, versie QQQ: De afdeling Onderzoek Informatie wil weten op welke manier de meerwaarde van de DAI het beste aangetoond kan worden. Bovendien wil ze een demonstrator laten ontwikkelen om bij de repository managers draagvlak te creëren. Ook wil ze aan de Europese collega's laten zien wat de voordelen zijn van de Nederlandse kennisinfrastructuur en de DAI.]

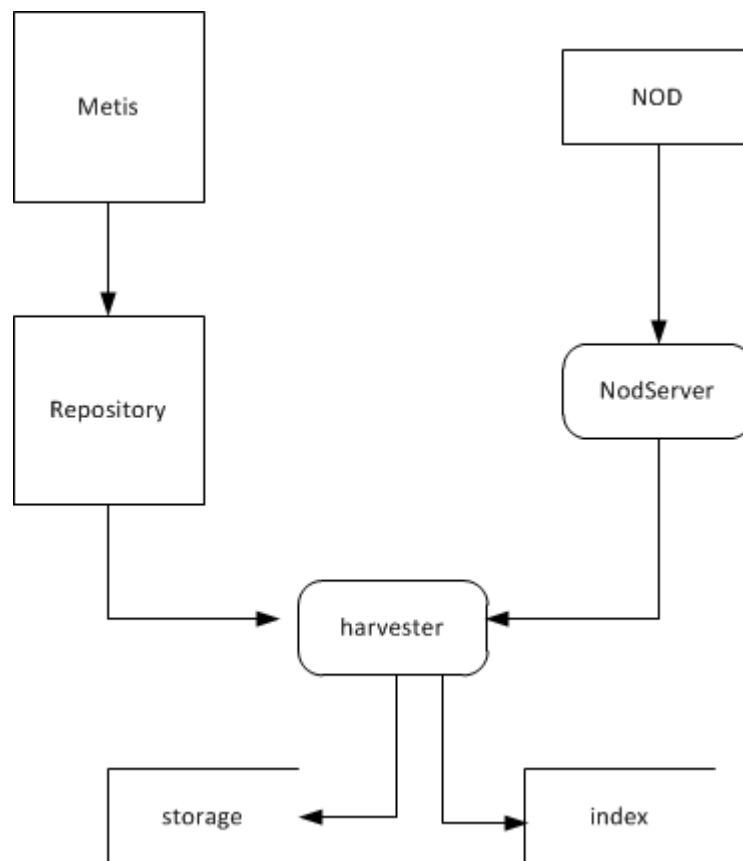
Hoe kunnen de mogelijkheden en de meerwaarde van de Digital Author Identification gedemonstreerd worden op een manier die de verschillende stakeholders aanspreekt?

2 Achtergrond

In de wetenschappelijke wereld worden onderzoekers beoordeeld op hun output. Om die te meten, wordt gebruik gemaakt van Metis¹ systemen, een Current Research Information System (CRIS) die alle Nederlandse universiteiten gebruiken. Repositories worden ingezet om de publicaties langdurig op te slaan. Hier wordt naast de metadata (titel, auteur, tijdschrift etc) ook de full text opgeslagen. In het geval van een open access publicatie is deze full text voor iedereen toegankelijk.

2.1 De dataflow

Voor het harvesten wordt gebruik gemaakt van Meresco, een Open Source product van CQ2. Deze haalt via het OAI-PMH protocol de metadata uit de repositories en indexeert deze. Ook wordt de metadata in een aantal formaten opgeslagen, waaronder het originele record.



Figuur 2: Gane Sarson DataFlowDiagram

Legenda van *Figuur 2: Gane Sarson DataFlowDiagram* :

Metis

Metis is de Nederlandse implementatie van een Current Research Information System

1 zie glossary

Repository

in de repositories wordt de metadata van publicaties en datasets opgeslagen

NOD (Nederlandse Onderzoek Databank)

Oracle DBMS

de data betreft onderzoekers, instellingen en onderzoekbeschrijvingen en de onderlinge relaties. zie ook de bijlage 8.5: *NOD Class Diagram*

storage

filesystem waarin elk record in de verschillende gewenste formaten wordt opgeslagen. Elk record wordt in ieder geval het originele formaat opgeslagen en in know_short, een eigen NARCIS formaat.

index

De index bevat alle data waarop gezocht kan worden. Net als bij *storage* is dit een versimpelde weergave van de werkelijkheid, aangezien er verschillende indexen zijn, bijvoorbeeld voor facetten, voor untokenized velden etc.

XML OAI met DC of DIDL/MODS

XML formaat van de Library of Congres om metadata uit te wisselen in antwoord op een OAI-PMH vraag.

zie ook bijlagen 8.3, 8.4 en 8.5.3

XML OAI met NOD XML

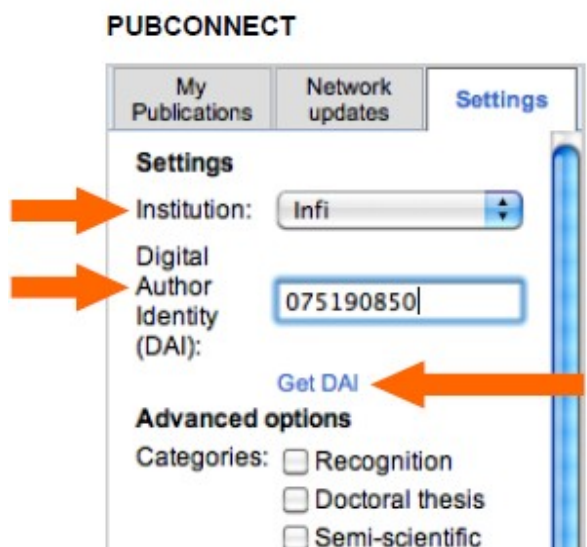
XML formaat van de Library of Congres om metadata uit te wisselen in antwoord op een OAI-PMH vraag.

zie ook bijlage *Error: Reference source not found*

2.1.1 Metis

Metis is de Nederlandse implementatie van een Current Research Information System (CRIS). Alle 13 Nederlandse universiteiten hebben een Metis, waarin metadata m.b.t. de onderzoekers, projecten en publicaties bewaard wordt.

De Metis beheerders zien het invoeren van de DAI niet als in hun belang. Om de DAI bij de de beheerders te promoten, heeft SURFfoundation begin 2010 de widget PubConnect laten ontwikkelen. Met PubConnect kan een onderzoeker zijn of haar publicatielijst, met de data zoals die is opgenomen in een Metis, presenteren op de eigen persoonlijke website. Dit wordt op basis van de DAI van die onderzoeker opgehaald. Daarnaast kan de onderzoeker een netwerk van mensen creëren rondom de publicatielijst. Een toepassing van Open Social en Google Friend Connect.



Figuur 3: DAI gebruikt in Pub Connect

2.1.2 NOD

De Nederlandse Onderzoek Databank is een openbare online databank met informatie over wetenschappelijk onderzoek, onderzoekers en onderzoeksinstituten (zie bijlage 8.5). De NOD bestrijkt alle wetenschappelijke disciplines (multidisciplinair) en is daarmee de nationale etalage voor onderzoekinformatie.

Doel van de NOD

- Toegankelijk en zichtbaar maken van Nederlands publiek gefinancierd onderzoek voor doelgroepen als wetenschappers, beleidsfunctionarissen, bedrijfsleven, media, wetenschappelijke uitgeverijen, informatiediensten en het brede publiek op nationale, Europese en mondiale schaal.
- Zichtbaar en vindbaar maken van kennis over onderzoekers (experts) en onderzoeksinstituten voor bovengenoemde doelgroepen op nationale, Europese en mondiale schaal (etalage voor wetenschappelijk onderzoek).

Een groot deel van de NOD bestaat uit feitelijke data. (duur van een onderzoek, titel, deelnemers etc.)

De NOD bevat daarnaast ook classificaties. De NOD Classificatie is een schema van codes waarmee experts en instellingen kunnen worden ingedeeld en teruggevonden (zie bijlage 8.5). De classificatie bestaat uit aandachtsgebieden (A) en wetenschappelijke disciplines (D). Per code worden 10 hoofdrubrieken gegeven met maximaal drie onderliggende niveaus. Daarnaast bevat de classificatie een C-code. Deze code omvat een zestal interdisciplinaire onderzoeksgebieden. De NOD Classificatie is in 2009 herzien.

Deze classificatie wordt handmatig toegekend. Zie ook *Tabel 5: aantal personen (totaal en met DAI) dat een classificatie term heeft toegekend gekregen*

De interface www.onderzoekinformatie.nl gaat 1 mei 2011 off line. Tegen die tijd is de functionaliteit opgenomen in de NARCIS Portal.

2.1.3 MODS

In eerste instantie werden de repositories geharvest in Dublin Core (DC). Dit formaat is verplicht wanneer het communicatie protocol OAI-PMH wordt gebruikt. Dublin Core heeft een aantal serieuze beperkingen. Daarom is het in Nederland opgevolgd door DIDL, een XML formaat met uitbreidingsmogelijkheden in de vorm van DIDL containers. Onder auspiciën van de Library of Congress is het MODS formaat vastgesteld, dat in een DIDL container verpakt kan worden. De huidige versie van MODS die in Nederland gebruikt wordt is 3.2. Niet elke repository werkt met dit formaat, sommige leveren in Dublin Core. Anderen gebruiken DIDL met miniMODS, een compromis-voorloper van MODS met alleen een uitgebreidere manier om persoonsinformatie bij een item op te slaan. Dit is voor dit project voldoende, aangezien hier ook de DAI's in zitten, als die voorhanden zijn. (zie *Figuur 7: namePart, gebruikt in MODS en miniMODS*)

2.2 Verrijkte Publicaties

SURFfoundation is begin 2011 gestart met een project over het inzichtelijk maken van de relaties tussen onderzoekdata, publicaties en datasets. Zij doen dit in de vorm van Verrijkte Publicaties. Een Verrijkte Publicatie is een aggregaat van samenhangende

objecten, waaronder tenminste één publicatie. Zo kan bijvoorbeeld een onderzoek met de bijbehorende dataset en de publicaties als één aggregaat worden gezien. Deze aggregaat worden gevat in een Resource Map. Deze wordt dan in RDF beschikbaar gesteld waardoor er met behulp van SPARQL query's beantwoord kunnen worden.

NARCIS vervult hierin de rol van nationale aggregator. Bovendien toont het de Resource Maps. Deze ReM zijn eilandjes, want ze zijn onderling niet verbonden. Wel kan binnen de ReM elk onderdeel als focuspunt gekozen worden en de bijbehorende metadata getoond worden.

Een mogelijke manier om de ReMs met elkaar te verbinden zou de DAI kunnen zijn. Ook de DOI, in Nederland gerealiseerd d.m.v. een URN, kan als verbinding fungeren.

Dit project is recent gestart, de voorgestelde periode is februari - oktober 2011. Zie voor de presentatie van de kick-off bijlage 8.10.

2.3 De invoering van de DAI

SURFfoundation heeft in begin deze eeuw met DAREnet een eerste aanzet gegeven om deze metadata te harvesten en via één loket aan te bieden. Om het vullen van de repositories met metadata te bevorderen, introduceerden ze de 'keur der wetenschap'. Dit bestond uit een select gezelschap van ongeveer 200 wetenschappers die een prominente plaats op de DAREnet-portal kreeg. Deze 'keurlingen' kregen ook een persoonlijke pagina, compleet met foto en bibliografie. Deze bibliografie werd opgesteld door in de index te zoeken naar alle publicaties waarbij een auteur dezelfde naam en voorletters had. Dit leverde vaak publicatielijsten op waarbij niet alle publicaties van de betreffende auteur waren. Het kwam ook voor dat publicaties van de auteur niet werden gevonden, omdat ze onder een andere schrijfwijze van de naam in de repositories stonden.

Om dit probleem te verhelpen is de Digital Author Identification ingevoerd, uitgegeven door OCLC. OCLC PICA² is een internationaal samenwerkingsverband tussen wetenschappelijke bibliotheken. Hier onderhoudt men een persoonsnamenthesaurus, de NTA. Alle auteurs die hierin worden opgenomen, krijgen automatisch een uniek nummer toegekend (een zogenoemde PPN). Dit nummer is later gebruikt als DAI en als zodanig opgenomen in de verschillende lokale onderzoeksregistratiesystemen, waaronder Metis.

Het is mogelijk dat een onderzoeker meerdere PPNs heeft. Bij de invoering van de DAI is ervoor gekozen om één PPN per onderzoeker te promoveren tot DAI. Hierbij heeft OCLC wat vergissingen begaan, door dezelfde DAI's meermalen uit te geven, of één en dezelfde onderzoeker meer dan één DAI toe te kennen. Deze vervuiling wordt langzaam maar zeker weer opgeschoond, maar is natuurlijk ook doorgesijpeld naar de verschillende repositories.

Ook de Nederlandse Onderzoek Databank (NOD) die de afdeling Onderzoek Informatie beheert, heeft een deel van de onderzoekers in de NOD gekoppeld aan een DAI. In *afbeelding 1* is aangegeven hoe de DAI gebruikt kan worden als koppeling tussen de repositories onderling en tussen de repositories en de NOD.

² OCLC PICA is een niet-commercieel samenwerkingsverband, anders dan Thomson Reuters, de beheerder van de ResearcherID.

3 Stakeholders en hun wensen

De vraag waar dit onderzoek een antwoord op zoekt, luidt:

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

Bij dit project zijn verschillende stakeholders betrokken. Niet alleen het MT van de universiteiten en de repository managers, maar natuurlijk ook de projectleiding van NARCIS.

3.1 Repository Managers

De Metis levert de metadata aan de repositories. De repositories vormen een elektronisch depot voor de wetenschappelijke output van het instituut. Hier wordt dus de metadata opgeslagen en de content. Als er geen embargo meer op rust, is ook de content publiek toegankelijk.

De repository manager is met name geïnteresseerd in een consistente repository, zichtbaarheid van de inhoud, juiste weergave en leuke downloadcijfers.

Repository managers zien vooral voordelen in het verbinden van de afzonderlijke repositories. Zij zien graag de volledige publicatielijsten van hun onderzoekers, ook wanneer de publicaties elders zijn verschenen. Ze zijn voornamelijk geïnteresseerd in overzichten waarin de onderzoeker centraal staat.

3.2 Management

Het management van de universiteiten en onderzoeksinstituten is met name geïnteresseerd in de zichtbaarheid van zijn onderzoekers en de correctheid van de gegevens. Bovendien moet de Return on Investment interessant zijn.

MT is gericht op de eigen onderzoek en onderzoekers, en wil dat de output van hun eigen onderzoekers onder de aandacht komt. Bijvoorbeeld: wat wordt er in een jaar gepubliceerd vanuit een bepaalde faculteit?

Welk onderzoek vindt er bij de faculteiten plaats, en op welke vlakken?

3.3 NARCIS projectleiding

NARCIS wil zich graag profileren als de nationale aggregator voor wetenschappelijke output. Hiertoe is het verrijken van de data door die met elkaar te verbinden een goed middel. Dit kan door de objecten naar elkaar te laten verwijzen, maar ook in een andere setting, dmv verrijkte publicaties.

3.4 Uitvoerder

De opdrachtnemer is als uitvoerder niet zozeer stakeholder van het eindresultaat, maar wel van het project.

De afstudeerder voegt zelf immers geen eisen toe aan het eisenpakket van de andere

stakeholders. Wel zijn er eisen en wensen met betrekking tot de ontwikkeling van het product binnen de looptijd van het afstudeerproject.

- Haalbaarheid in tijd en doorlooptijd.
- Haalbaarheid qua beschikbare data.
- Haalbaarheid qua capaciteiten.

3.5 wensenlijst

Uit de gesprekken (zie bijlage 8.8) heb ik de volgende wensenlijst gedestilleerd:

- C1. hoog dai gehalte
de rol van de DAI in de potentiële toepassing. waar ligt het accent? Leveren meer DAI's een beter resultaat op?
- C2. Onderling verbinden van repositories
relaties leggen tussen de publicaties en de datasets van de verschillende repositories. Naarmate er meer relaties tussen de repositories gelegd kunnen worden, komt er een completer beeld van het werk van een onderzoeker.
- 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 niet alle data die beschikbaar is, ook gevonden wordt. Dit gaat niet over de inherente incompleteheid 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 bij deze toepassing?
- C8. auteur staat centraal
zorgt deze toepassing 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.
- C10. Instituut staat centraal
zorgt deze toepassing ervoor dat de informatie voor een specifiek instituut beter tot zijn recht komt?

3.6 De wensen gewogen door de stakeholders

De verschillende stakeholders waarderen deze wensen op hun eigen manier. Onderstaande tabel toont hoe de wensen t.o.v. elkaar gewogen zijn: is de horizontale wens belangrijker dan de verticale, dan krijgt de horizontale wens 1 punt. De laatste kolom toont de som voor deze wens, het uitgangspunt voor de weegfactor.

	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 1: matrix waarin de wensen paarsgewijs beoordeeld worden

Dit levert voor de drie stakeholders management (MT), repository managers (rm) en de NARCIS projectleiding de volgende weegfactoren³:

	mt	r m	NARCIS
C1	0	0	6
C2	1	4	6
C3	8	9	9
C4	7	5	4
C5	3	5	3
C6	4	2	4
C7	2	2	5
C8	5	7	1
C9	9	8	7
C10	6	2	0

Tabel 2: de weegfactoren voor de drie stakeholders

³ zie voor verder cijfermateriaal bijlage 8.7

4 Potentiële toepassingen

4.1 Mogelijkheden en beperkingen

Er kan niet meer getoond worden dan wat er beschikbaar is. Dit betekent dat NARCIS niet de indruk moet wekken volledig te zijn. Wel kan correctheid worden nagestreefd: de data die getoond wordt, de relaties die gelegd worden zijn in overeenstemming met de aanwezige data.

Sterker nog: een omissie in de DAI-data die duidelijk wordt door de gekozen toepassing is juist een doel van het project: een incentive om die data aan te (laten) vullen.

De data is opgeslagen als files op een filesysteem, dus niet in een klassieke database. De index kan bevraagd worden door middel van SRU met CQL. Alleen de velden die geïndexeerd zijn, zijn beschikbaar. Dit kan uitgebreid worden door het DNA van de indexer aan te passen en een re-index uit te voeren, een tijdrovende bezigheid. Op dit moment duurt dit zo'n 30 uur, en kan alleen plaatsvinden op daluren, aangezien het de zoekfunctie ernstig vertraagt. De doorlooptijd is dus een aantal dagen.

vragen als *hoeveel publicaties met DAI zitten in de index?* zijn via CQL wel te stellen. Maar *hoeveel verschillende DAI zitten in de repositories?* is via CQL niet te vragen.

4.2 Lijst van potentiële toepassingen

In februari is in de setting van een brainstorm gesproken over de mogelijkheden van de DAI. De ideeën die hier geopperd werden, herformuleerd en samengebracht, heb ik in gesprekken aan derden voorgelegd, waardoor er aanvullingen en verbeteringen konden worden aangebracht. Het resultaat hiervan is de volgende lijst van potentiële toepassingen.

4.2.1 Woordenwolk samenstellen bij persoon

Een woordenwolk is een manier om de karakteristieke termen bij een onderwerp weer te geven. Op de persoonspagina staat nu een woordenwolk die is samengesteld op basis van de termen die een onderzoeker zelf opgeeft bij zijn expertise. Deze zou uitgebreid kunnen worden aan de hand van overige data uit de NOD, en ook publicaties die op basis van de DAI aan de persoon gekoppeld worden. Het automatisch vaststellen van de belangrijkste termen is geen triviale zaak. Dit geldt met name als de tekst uit weinig woorden bestaat.

De nadruk ligt hier meer op tekstprocessing dan op de meerwaarde van de DAI. Het invoeren van meer publicaties met DAI levert niet noodzakelijk een betere woordenwolk op. Bij slechte prestatie van het verwerken van de tekst kan een verkeerde voorstelling van zaken gegeven worden. Wanneer het goed gedaan wordt, levert het in een oogopslag de belangrijkste termen bij een persoon.

4.2.2 Extra bronnen raadplegen

Naast de repository van de eigen organisatie, heeft een onderzoeker vaak ook op andere plaatsen publicaties gedeponneerd, bijvoorbeeld bij PubMed. Ter completering van de

gegevens in NARCIS kunnen deze en andere archieven afgezocht worden naar publicaties / datasets die niet in de geharveste repositories aanwezig zijn. Dit kan op basis van DAI (als die aanwezig is) of op basis van gegevens in de NOD, zoals naam, instituut, expertise. De publicaties die op deze manier gevonden worden, kunnen van een andere auteur zijn. Het is ook denkbaar dat er dubbelingen in de index komen, met de perikelen van het 'ontdubbelen'.

Voor de volledigheid vermeld ik hier een optie die vaak naar voren komt. Het bieden van de mogelijkheid aan individuele onderzoekers om hun eigen publicaties te uploaden. Dit vereist een nieuwe infrastructuur, eigenlijk een eigen repository. Een manier om vast te stellen of de publicatie al bekend is, en natuurlijk een identificatie en autorisatie mechanisme. Dit valt echter buiten de scope van deze opdracht, aangezien dit geen link heeft met de DAI. Bij het uploaden van een publicatie zou de metadata gecreeerd worden in plaats van geharvest.

4.2.3 Experts vinden

Het vinden van relevante experts, uitgaande van een publicatie of andere tekst. Wanneer het om personen gaat die ook in de NOD staan, is het mogelijk een academische titel (UHD, hoogleraar) af te dwingen.

Het eigenlijke probleem hier is het definiëren van een afstandsbeleg tussen profielen van personen en samenvattingen van teksten.

In de NOD is van zo'n vijftienduizend personen de expertise bekend. Hiernaast kunnen de publicaties van de personen gebruikt worden. De rol van de DAI binnen deze toepassing is gelimiteerd tot het vinden van deze publicaties. Dit is maar een zeer klein deel van de totale toepassing.

CON: Vaststellen van expertise op basis van publicaties is net als een woordenwolk een niet-triviale taak.

4.2.4 Profielpagina's koppelen

Onderzoekers hebben op meer plaatsen profielpagina's, zoals bijvoorbeeld COS of LinkedIn. Probeer deze informatie te koppelen, en als de koppeling tot stand is gekomen te integreren. Hierbij moeten privacy overwegingen meegenomen worden.

4.2.5 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.

4.2.6 Academische loopbaan tonen

De academische verdiensten van een persoon tonen, aan de hand van z'n publicaties, werkverbanden en onderzoek. Misschien met een tool als <http://www.dipity.com/>. Data

komt uit repositories in de vorm van publicaties en datasets en uit de NOD in de vorm van onderzoek.

Data kan ook van buiten komen, bijvoorbeeld uit de GGC⁴ van OCLC-PICA, de uitgever van de DAI. In de GGC zijn bij werkverbanden van onderzoekers de begindata opgeslagen, zie bijlage 8.6

. In de NOD wordt alleen bijgehouden waar een persoon op dit moment werkzaam is, zonder begin- of einddatum.

De NOD is opgezet als databank voor lopend onderzoek. Hierin was de historie niet van belang. Hoewel deze gedachte in het licht van persistente opslag steeds meer terrein verliest, is dit voorlopig voor een tijdslijn niet voldoende.

De GGC heeft een privacy statement uitgegeven (zie Privacy Statement GGC). Hierin beschrijven ze het doel van de GGC, namelijk het identificeren van auteurs. De data mag niet voor een ander doel gebruikt worden.



Figuur 4: dipity timeline

4.2.7 Onderzoek aan publicatie koppelen

Het idee is om onderzoekers uit de NOD te koppelen aan publicaties en datasets.

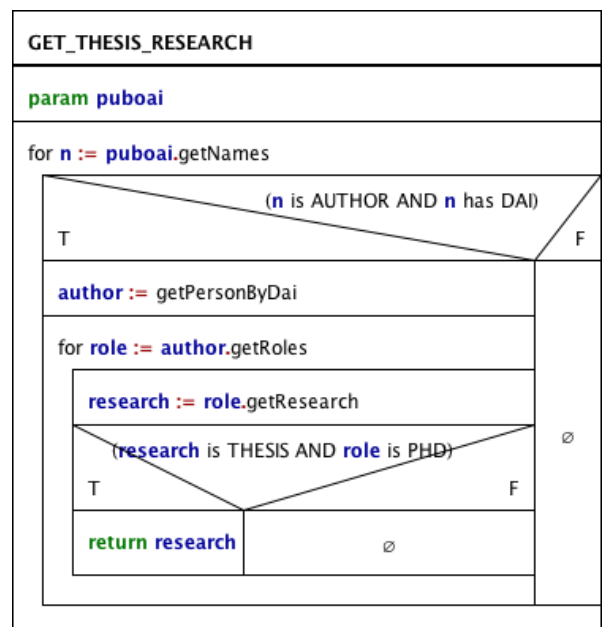
Dit is een lastige bezigheid, er moet gematched worden op titel, samenvatting, auteur, etc. Bijvoorbeeld: onderzoekers die gekoppeld zijn aan een onderzoek in de NOD, en samen een publicatie hebben ten tijde van dat onderzoek. Dit kan een mismatch opleveren.

Een manier om de precision (het aantal terecht gevonden publicaties gedeeld door het totaal aantal gevonden publicaties, zie verder bijlage 8.11) te verhogen is het verkleinen van de pool van publicaties tot alleen doctoral theses.

Promotieonderzoek heeft vaak dezelfde titel als de bijbehorende thesis. Om vast te stellen of een onderzoek een promotie-onderzoek betreft, kan gekeken worden naar het attribuut isThesis. Bij de werkzaamheid wordt bijgehouden in welke rol een persoon bij een onderzoek betrokken is. Wanneer dit als promovendus of promotor is, kan worden aangenomen dat het onderzoek een promotie-onderzoek is. Dit geldt voor zo'n vijftienduizend onderzoeken.

Wanneer de DAI van de promovendus gelijk is aan de DAI van de auteur van het proefschrift, kan aangenomen worden dat die bij elkaar horen. (Zie ook Figuur 5)

Dit zou goed kunnen worden uitgevoerd in het kader van het VPN-project. Niet slechts



Figuur 5: PSD voor het identificeren van het thesis onderzoek bij een gegeven publicatie

⁴ Zie de glossary

de Verrijkte Publicaties die aangeleverd worden harvesten en indexeren, maar ook zelf VP's genereren, en aanbieden in RDF.

4.2.8 Publicaties bij instituten tonen

Tonen van de publicaties per universiteit / faculteit, ook als de publicatie bij een ander instituut in de repository zit.

Op dit moment wordt bij gebrek aan beter de repository waar een publicatie is gedeponeerd beschouwd als de instelling waar het onderzoek en dergelijke ook heeft plaatsgevonden. Dit is soms te kort door de bocht, bijvoorbeeld wanneer onderzoekers van meerdere instituten gezamenlijk een publicatie schrijven.

Wanneer de auteur ook in de NOD voorkomt, is bekend bij welk instituut hij op dit moment werkzaam is. Dit is echter geen garantie dat het werk onder de auspiciën van dat instituut is gedaan.

CON: Historie van werkverbanden wordt niet bewaard.

CON: kan onjuiste informatie opleveren

4.2.9 Emerging topics identificeren

Per NOD-classificatie de bijbehorende publicaties identificeren. (personen hebben een classificatie, deze personen hebben een DAI, bij de DAI horen publicaties. Aannee: alle publicaties van een persoon hebben betrekking op deze classificatie) Deze groeperen per jaar. Hieruit zijn dan de stijgers en dalers te bepalen. Zie bijvoorbeeld [NRC01]

4.2.10 Overlappen tussen onderzoeksgebieden vaststellen

Welke onderzoeksgebieden kruipen naar elkaar toe? Van publicaties vaststellen in welke onderzoeksgebieden ze vallen. Dit zou kunnen door te kijken naar de faculteit van de auteur. Ook de classificaties van de auteur zijn hiervoor te gebruiken. Elfduizend onderzoekers in de NOD hebben een DAI en tenminste 1 classificatie term. Zie ook *Figuur 10: de persoon in de NOD met de DAI en classificatie termen*.

Wanneer dit voor een aantal opeenvolgende tijdsperiodes gedaan wordt (bijvoorbeeld alle publicaties van een jaar) naar elkaar toe groeiende gebieden onderkend worden. Dit is een uitbreiding op 4.2.6: Academische loopbaan tonen.

4.2.11 Verschuivende interesse-gebieden tonen

Chronologisch het verloop van interesse-gebieden voor een persoon tonen. Een interessante groep hiervoor zijn de voormalige keurlingen. Dit zijn vaak de belangrijkste onderzoekers voor een instituut, van wie alle publicaties in de repositories zitten, die expertise hebben etc. Dit is een uitbreiding op Academische loopbaan tonen, aangezien hierbij de onderwerpen van de publicaties bepaald worden, en die in een tijdslijn uitgezet worden. De interesse-gebieden eruit halen is een meer tekst georiënteerd proces. De relatie met de DAI is niet eenduidig: meer DAI's leveren niet noodzakelijk betere interesse lijnen op.

4.3 Potentiële toepassingen volgens de wensen

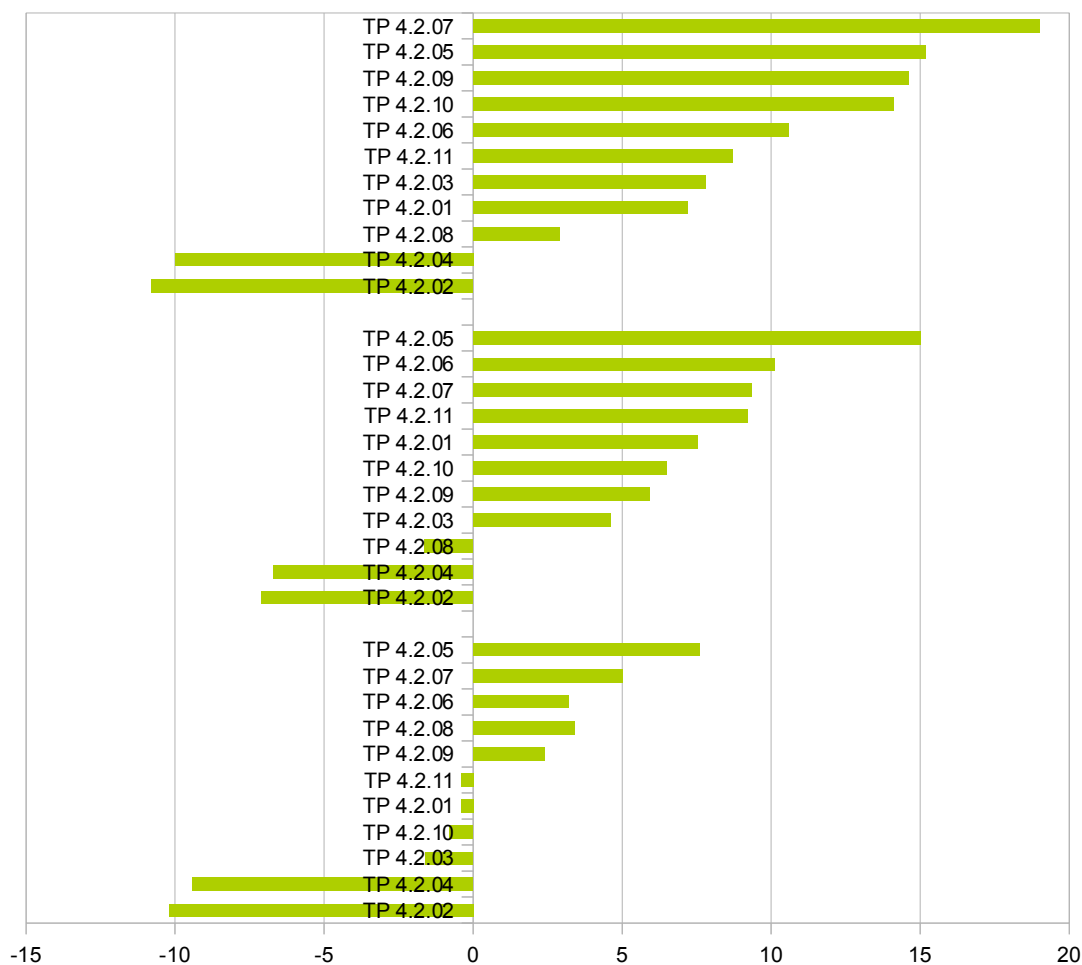
Om een keuze te maken uit deze potentiële toepassingen, heb ik ze gescoord aan de hand van de genoemde 9 wensen. Deze score is besproken met de bedrijfsmentor en met drs. Elly Dijk, de voormalig voorzitter van de WRM en lid van de Narcis Advies Raad.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
4.2.01	50	1	60	2	2	70	1	1	0	0
4.2.02	10	0	70	0	0	40	0,5	1	1	0
4.2.03	10	1	40	2	3	40	1	0	0	0
4.2.04	10	0	70	0	0	20	0,5	1	1	0
4.2.05	100	1	0	1	3	0	0	1	0	0
4.2.06	100	1	10	0	3	0	0,5	1	1	0,5
4.2.07	100	1	0	0	2	0	1	0	0	0
4.2.08	50	0	50	0	1	40	1	0	0	1
4.2.09	100	1	0	2	2	70	1	0	0	0
4.2.10	100	1	30	2	3	40	1	0	0	0,5
4.2.11	100	1	60	2	3	70	0,5	1	0	0

Tabel 3: de potentiële toepassingen gescoord op de wensen

5 De stakeholders en de potentiële toepassingen

wanneer de scores uit *Tabel 3: de potentiële toepassingen gescoord op de wensen* verrekend worden met de weegfactor uit *Tabel 2: de weegfactoren voor de drie stakeholders*, krijgt elke toepassing voor elke stakeholder een eigen score.



Figuur 6: de volgorde voor het management, de repository managers en de NARCIS projectleiding

Figuur 6 toont de respectievelijke volgorden voor de stakeholders MT, repository managers en de NARCIS projectleiding.

6 Conclusie

Bij alle drie staan de toepassingen *Onderzoek aan publicatie koppelen* en *Coworkers tonen* in de top drie. Coworker heeft een sterke nadruk op de onderzoeker. De professionele omgeving wordt getoond. Promotie-onderzoek daarentegen legt de nadruk op een publicatie en geeft daarbij de omgeving. Met deze twee toepassingen wordt de hele reikwijdte van de DAI bestreken.

7 Literatuur

Literatuur

NRC01: Warna Oosterbaan, Dit doen we goed – scival laat zien wie uitblinkt in onderzoek , 2010

- Research Information Systems, the Dutch way
Marjan Vernooy-Gerritsen, SURFfoundation
Bert van Zomeren, Delft University of Technology
<https://blackboard.hhs.nl/bbcswebdav/users/10102515/ondersteuning/van%20zomeren.odp>
- <http://dublincore.org/documents/dces/> [accessed 2011-02-16, zotero attachment 14]
- <http://www.surfoundation.nl/nl/projecten/Pages/Metiswidget.aspx> [accessed 2011-02-16, zotero attachment 12]
- <http://science.thomsonreuters.com/press/2008/8429910/> [zotero 16]
- RAPPORTAGE PROJECT DAI-UITROL,
https://blackboard.hhs.nl/bbcswebdav/users/10102515/ondersteuning/Rapportage_voor_WISH.doc

8 Bijlagen

8.1 Cijfers huidige stand van zaken m.b.t. de DAI

ook interessant als baseline voor een latere vergelijking

1. hoeveel verschillende DAI s zitten in de repositories?

Naam	Totaal aantal dai-auteurs (28-11-2007)
Rijksuniversiteit Groningen	2189
Universiteit Twente	1196
Universiteit van Amsterdam	4654
Erasmus Universiteit Rotterdam	3013
Radboud Universiteit Nijmegen	3911
Technische Universiteit Delft	4539
Technische Universiteit Eindhoven	2397
Universiteit Leiden	
Universiteit Maastricht	2654
Universiteit Utrecht	5259
Universiteit Tilburg	
Vrije Universiteit Amsterdam	2779
Wageningen Universiteit & Researchcentrum	4549
Centrum voor Wiskunde en Informatica (CWI)	233
KNAW	1033

Tabel 4: cijfers tijdens de uitrol van de DAI van de 15 deelnemende instituten. uit: RAPPORTAGE PROJECT DAI-UITROL

[de volgende waarden zijn gemeten op 23 februari 2011]

2. hoeveel publicaties (procent) hebben tenminste één DAI?

235.427 van de 618.266 = 38 %

3. hoeveel DAI s zitten er in de NOD?

18.949

4. hoeveel DAI s in de NOD hebben een classificatie?

aantal classificaties	aantal personen	aantal dai's	
tenminste 1	18831	11759	62,1%
2	12881	8388	44,3%
3	5646	3749	19,8%
4	2067	1392	7,3%
5	645	461	2,4%
6	193	127	0,7%

Tabel 5: aantal personen (totaal en met DAI) dat een classificatie term heeft toegekend gekregen

zie ook *Figuur 9: de associaties tussen persoon, organisatie en onderzoek en de classificatie termen*

5. hoeveel DAI s in de NOD hebben een C classificatie?

1031 (1657 personen met/zonder dai)
er zijn 6 C-classificaties: C10000 tot C60000

8.2 Verklarende woordenlijst

Zie de [glossary]

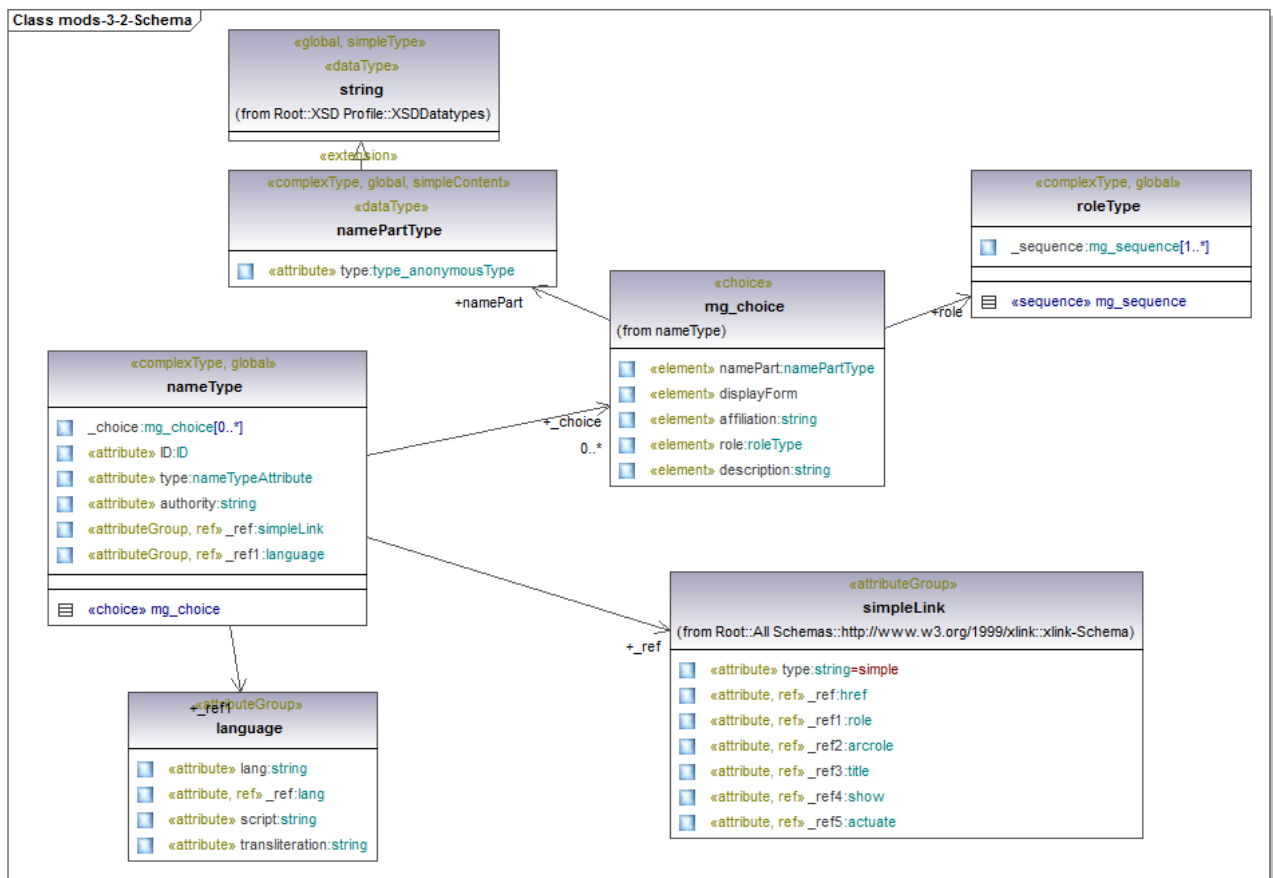
8.3 DC

De Dublin Core Metadata Element Set is een set van vijftien attributen die gebruikt kunnen worden om bibliografische objecten te beschrijven. 'Dublin' wegens de plaats waar de workshop in 1995 is gehouden in Dublin, Ohio. 'Core' omdat de attributen breed en generiek zijn, geschikt voor velerlei typen objecten.

De vijftien attributen zijn onderdeel van een grotere set metadata vocabularies en technische specificaties onderhouden door de Dublin Core Metadata Initiative (DCMI).

8.4 DIDL / MODS XSD

Didl is een xml formaat om complexe digitale objecten te beschrijven. Dit gebeurt aan de hand van concepten als Containers, Resources, Components en Items. Hierbinnen kan door middel van MODS (zoals gedefinieerd door de Library of Congress) de metadata van de objecten beschreven worden. In figuur 7 is de NamePart gemodelleerd, een deel van de MODS die door de repositories gebruikt wordt.



Generated by UModel

www.altova.com

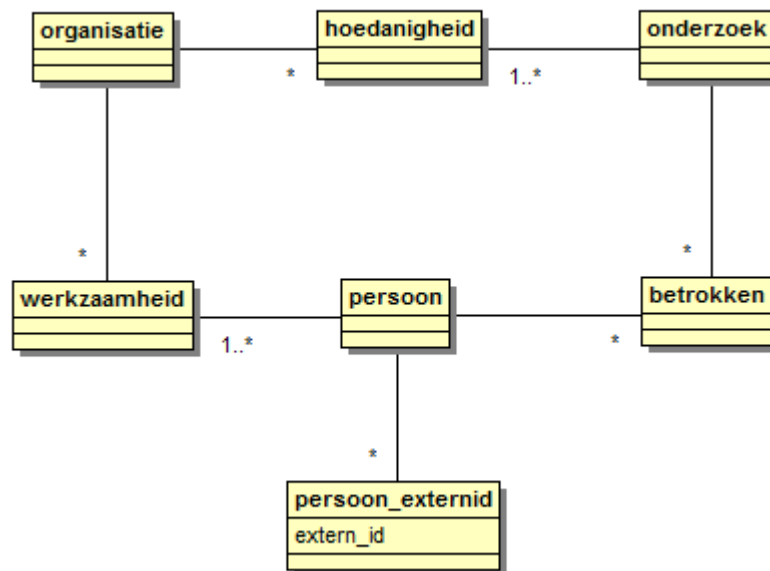
Figuur 7: namePart, gebruikt in MODS en miniMODS

8.5 NOD Class Diagram

De NOD database bestaat uit zo'n veertig tabellen. Figuur 8 toont de belangrijkste concepten: organisaties, onderzoeken en personen. Deze zijn met elkaar verbonden via tussentabellen die aangeven welke rol de relatie vervult. Bovendien wordt getoond hoe de DAI in de NOD verwerkt zit, namelijk als een van de externe identificaties die aan een persoon gekoppeld kunnen worden.

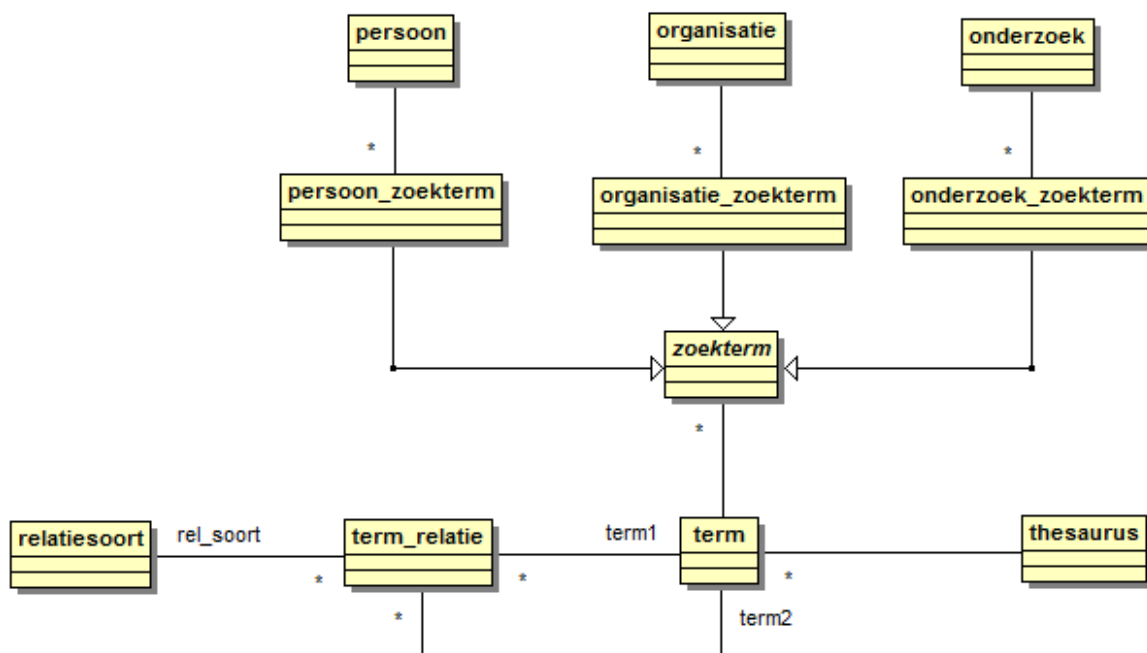
Figuur 9 toont de manier waarop classificaties, hier zoektermen genoemd, aan de personen, onderzoeken en organisaties gehangen kunnen worden. De classificaties zelf maken deel uit van een thesaurus en vormen onderling een boomstructuur.

8.5.1 Hoofdklassen



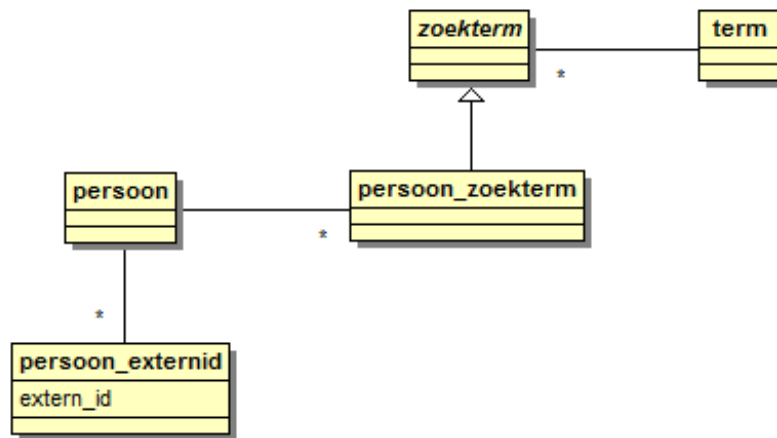
Figuur 8: domein model hoofdklassen NOD, waarin ook de DAI (als externid) aangegeven is

8.5.2 Classificatie



Figuur 9: de associaties tussen persoon, organisatie en onderzoek en de classificatie termen

8.5.3 DAI

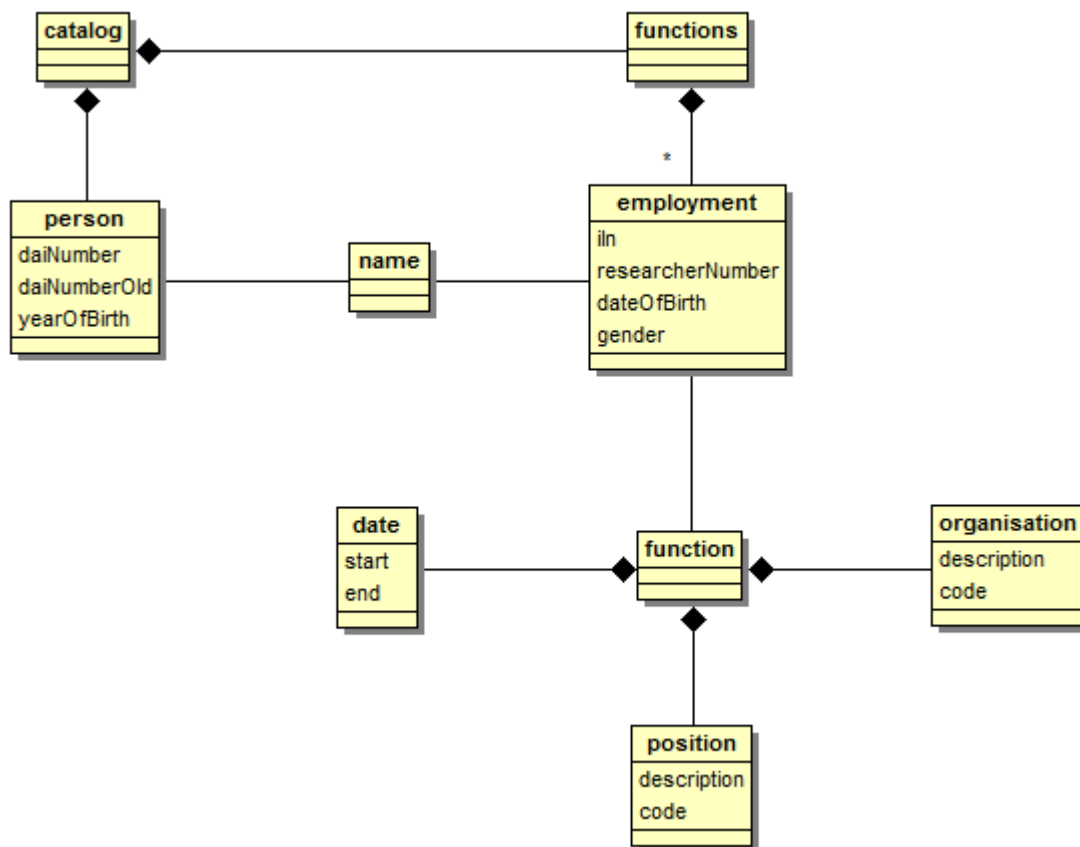


Figuur 10: de persoon in de NOD met de DAI en classificatie termen

8.6 PICA's DAI XML

Pica heeft een XML gedefinieerd die zij gebruikt om de resultaten van een zoekvraag terug te geven. De root-node is de catalog, waarin een persoon en een functions object zitten. De persoon kan twee DAI nummers hebben, een 'echte' en een 'oude'. In de praktijk hebben personen soms wel meer dan twee DAIs gehad. Deze zijn niet meer terug te vinden in de XML.

Aangezien elke DAI eerst een PPN was, zijn ze op die manier nog wel na te speuren, maar niet automatisch.



Figuur 11: DAI XML export van OCLC PICA

8.7 berekening weegfactoren criteria

Zie [TJR02]

8.8 gespreksverslagen

Zie algemene bijlage 4

8.9 Privacy Statement GGC

https://blackboard.hhs.nl/bbcswebdav/users/10102515/ondersteuning/NTA_Privacy_statement_20070103.doc

8.10 Kick-off Verrijkte Publicaties

https://blackboard.hhs.nl/bbcswebdav/users/10102515/ondersteuning/surffoundation_VP_Kick%20off%2019-02-2011.odp

8.11 Precision en Recall

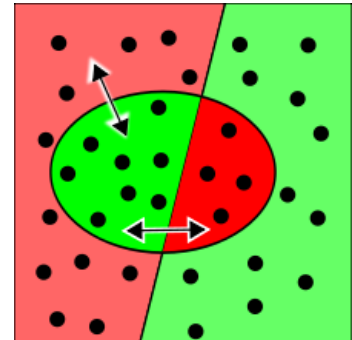
Precision en recall zijn termen uit de data mining en zijn als volgt gedefinieerd:

Precision: het aantal terecht gevonden documenten, gedeeld door het totaal aantal gevonden documenten.

In figuur 12 zijn de bolletjes links de relevante documenten, rechts de niet-relevante documenten. De bolletjes in de ovaal zijn de gevonden documenten. De groene bolletjes in de ovaal representeren de terecht gevonden documenten. Precision betekent dus de groene bolletjes in de ovaal gedeeld door alle bolletjes in de ovaal.

Recall: het aantal terecht gevonden documenten, gedeeld door het aantal documenten dat gevonden had *moeten* worden.

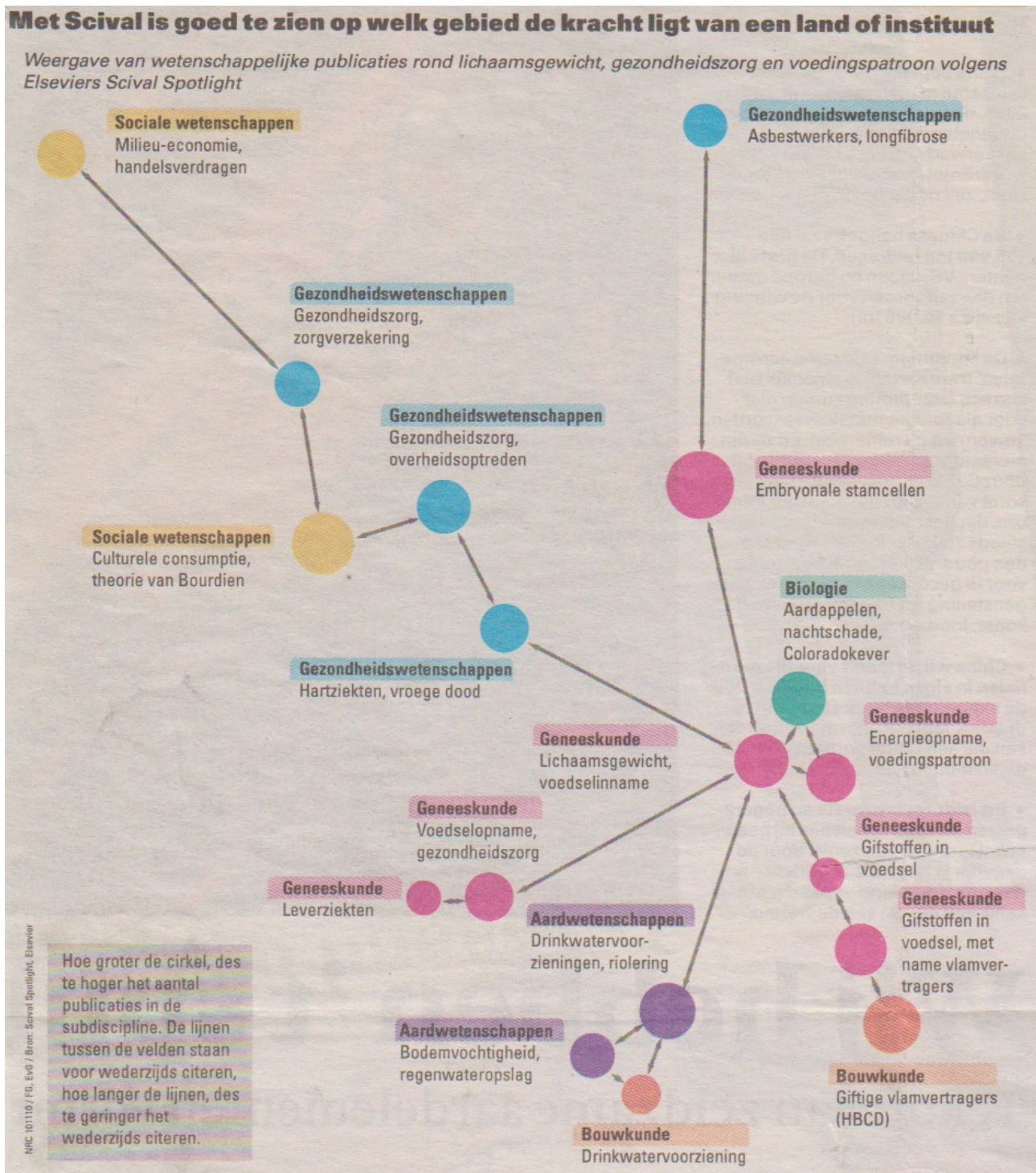
In termen van figuur 12: de groene bolletjes in de ovaal gedeeld door de rode bolletjes buiten de ovaal.



Figuur 12: precision en recall

8.12NRC: Dit doen we goed

[uit: nrc next: woensdag 10 november 2010, wetenschap]



Dit doen we goed

Scival laat zien wie uitblinkt in onderzoek

► Aan de hand van onderlinge citaten construeert Scival Spotlight van Elsevier onderzoeksclusters.

► Zo kun je zien waarin welk land precies goed is.

Door WARNA OOSTERBAAN

DELFT. Misschien weten de hoogleraren Piersma, Slob (beiden RIVM) en Savenije (TU Delft) het nog niet, maar zij en nog een heel stel andere onderzoekers dragen met hun activiteiten bij aan een onderzoeksgebied dat een van de wetenschappelijke speerpunten van Nederland is. Toch is het een onderzoeksgebied waar geen leerstoel voor bestaat, dat verspreid ligt over toxicologie, sociologie, aardwetenschappen, farmacologie en de studie van beroepsrisico's. Gemeenschappelijke thema's zijn lichaamsgewicht, gezondheidszorg en voedingspatroon. Het gebied bestaat dankzij het feit dat de onderzoekers een elkaar citerende, interdisciplinaire gemeenschap vormen. Blijkbaar profiteren ze van elkaars inzichten. En dat doen ze goed, want ze namen in dit gebied bijna 16 procent van de wereldproductie in de periode 2004-2009 voor hun rekening: 159 van de 1.018 wetenschappelijke artikelen die in de gehele wereld op dit terrein werden geschreven.

Dat kom je te weten als je Scival Spotlight opstart, een onlineproduct van Elsevier waarvan eergisteren een landenmodule op de markt kwam. Het is een loot aan de bibliometrische stam, de hulpwetenschap die de mon-

diale wetenschappelijke productiviteit pogt te meten. Basismateriaal is Scopus, een database van 45 miljoen wetenschappelijke artikelen.

Scival is een kistje met meetgereedschap voor iedereen die wil weten waar de kracht van een instituut of universiteit ligt en welke onderzoekers excelleren. Met de nieuwe landenmodule, 'Country Maps', worden allerlei metingen op nationaal niveau mogelijk. Zo leren we dat in de afgelopen vijf jaar onderzoekers in dienst van Nederlandse instellingen ruim 178.000 artikelen produceerden: ruim 2,2 procent van de wereldproductie van bijna 8 miljoen artikelen. Niet gek als je bedenkt dat de Neder-

► Nederland scoort goed op euthanasie, vliegtuiglawaai en ziekteverlof ◀

landers slechts 2 promille van de wereldbevolking vormen. In Groot-Brittannië wonen ongeveer 3,7 keer zoveel mensen, maar uit de Scival-database blijkt dat ze maar 3,4 keer zoveel artikelen produceren. En er zijn 18,7 keer zoveel Amerikanen, maar ze produceren maar 12,9 keer zoveel artikelen. Dat is wel meteen bijna 30 procent van de wereldproductie.

Het interessantste aan Scival Spotlight zijn de clusters van verwante onderzoeksgebieden die ermee kunnen worden geconstrueerd. „Moderne wetenschapsbeoefening houdt zich steeds minder aan de gebruikelijke indeling in disciplines”, zegt Michiel

Kolman, *senior vice president global academic relations* bij Elsevier. „Daarom richten wij ons op die multidisciplinaire clusters.”

Een cluster heet binnen Scival een competentie als de productie in artikelen binnen dat gebied een bepaalde drempelwaarde overschrijdt. Per land of instituut kan worden bepaald of dat zich op bepaalde clusters onderscheidt. Het moet dan binnen dat cluster een aanzienlijk marktaandeel hebben, de publicaties moeten vaak geciteerd worden en de citaties in de artikelen zelf moeten betrekking hebben op recente publicaties – een aanwijzing voor innovatieve kracht.

Met die criteria komt Scival voor Nederland tot 97 competenties, onderzoeksgebieden waar dit land het goed doet. Veel is biomedisch: heel goed scoort een cluster met trefwoorden ziekteverlof, chronische pijn en pijnintensiteit. Aan Nederlandse instellingen verbonden onderzoekers publiceerden meer dan 18 procent van de wereldproductie in dit cluster.

Ook doet Nederland het heel goed in een cluster met trefwoorden vliegtuiglawaai en geluidshinder. Ruim 13 procent van de wetenschappelijke wereldproductie op dit gebied komt uit Nederland en prof. Frewer van Wageningen Universiteit is met 14 publicaties de topauteur. Of neem het cluster palliatieve zorg, euthanasie en stervensbegeleiding: 18 procent.

Er zijn ook andere instrumenten op de markt, zoals Web of Science van concurrent Thomson Reuters. Maar daarmee kunnen niet dit type clusteranalyses worden uitgevoerd, zegt Kolman. Bovendien bevat Scival Spotlight circa 18.000 tijdschriften en Web of Science maar zo'n 10.000.

Weegfactoren

De verschillende stakeholders hebben de criteria C1 t/m C10 paarsgewijs vergeleken en hun voorkeur uitgesproken. Tabel 1, 2 en 3 tonen deze voorkeuren.

	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 1: voorkeuren NARCIS projectleiding

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	totaal
C1	-	0	0	0	0	0	0	0	0	0	0
C2	1	-	0	0	0	1	1	0	0	1	4
C3	1	1	-	1	1	1	1	1	1	1	9
C4	1	1	0	-	1	1	1	0	0	1	6
C5	1	1	0	0	-	1	1	0	0	1	5
C6	1	0	0	0	0	-	0	0	0	1	2
C7	1	0	0	0	0	1	-	0	0	0	2
C8	1	1	0	1	1	1	1	-	0	1	7
C9	1	1	0	1	1	1	1	1	-	1	8
C10	1	0	0	0	0	0	1	0	0	-	2

Tabel 2: voorkeuren repository managers (rm)

	C1	C2	C3	C4	C5	C6	C7	C8	C9	10	totaal
C1	-	0	0	0	0	0	0	0	0	0	0
C2	1	-	0	0	0	0	0	0	0	0	1
C3	1	1	-	1	1	1	1	1	0	1	8
C4	1	1	0	-	1	1	1	1	0	1	7
C5	1	1	0	0	-	0	1	0	0	0	3
C6	1	1	0	0	1	-	1	0	0	0	4
C7	1	1	0	0	0	0	-	0	0	0	2
C8	1	1	0	0	1	1	1	-	0	0	5
C9	1	1	1	1	1	1	1	1	-	1	9
C10	1	1	0	0	1	1	1	1	0	-	6

Tabel 3: voorkeuren beleidsmakers (bm)

De weegfactoren $w_{(C)}$ van de criteria voor de verschillende stakeholders in een tabel:

	NARCIS	b m	r m
C1	6	0	0
C2	6	1	4
C3	9	8	9
C4	4	7	5
C5	3	3	5
C6	4	4	2
C7	5	2	2
C8	1	5	7
C9	7	9	8
C10	0	6	2

Tabel 4: de weegfactoren van de drie stakeholders

De alternatieven zijn op deze criteria gescoord. De scores $S_{(A,C)}$ die de uitkomst van dit proces waren, zijn weergegeven in tabel 5.

titel		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
woordenwolk persoon	A1	50	1	60	100	66	70	100	1	0	0
extra bronnen	A2	10	0	70	0	0	40	50	1	1	0
expertfinder	A3	10	1	40	100	100	40	100	0	0	0
profielpagina	A4	10	0	70	0	0	20	50	1	1	0
coworker	A5	100	1	0	50	100	0	0	1	0	0
loopbaan	A6	100	1	10	0	100	0	50	1	1	50
promotieonderzoek	A7	100	1	0	0	66	0	100	0	0	0
koppeling instituut publica	A8	50	0	50	0	33	40	100	0	0	100
emerging topic	A9	100	1	0	100	66	70	100	0	0	0
overlap	A10	100	1	30	100	100	40	100	0	0	50
interesse verloop	A11	100	1	60	100	100	70	50	1	0	0

Tabel 5: de alternatieven gescoord op de criteria

De weegfactoren van de stakeholders zijn vermenigvuldigd met deze score en vervolgens opgeteld. Dit levert volgens de NARCIS projectleiding voor de alternatieven de waarden op getoond in tabel 6.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	
	100	1	100	100	100	100	100	1	1	100	
	6	6	9	4	3	4	5	1	7	0	
	1	1	-1	-1	1	-1	1	1	-1	1	gewicht
A2	0,6	0	-6,3	0	0	-1,6	2,5	1	-7	0	-10,8
A4	0,6	0	-6,3	0	0	-0,8	2,5	1	-7	0	-10
A8	3	0	-4,5	0	0,99	-1,6	5	0	0	0	2,89
A1	3	6	-5,4	-4	1,98	-2,8	5	1	0	0	4,78
A3	0,6	6	-3,6	-4	3	-1,6	5	0	0	0	5,4
A11	6	6	-5,4	-4	3	-2,8	2,5	1	0	0	6,3
A6	6	6	-0,9	0	3	0	2,5	1	-7	0	10,6
A10	6	6	-2,7	-4	3	-1,6	5	0	0	0	11,7
A9	6	6	0	-4	1,98	-2,8	5	0	0	0	12,18
A5	6	6	0	-2	3	0	0	1	0	0	14
A7	6	6	0	0	1,98	0	5	0	0	0	18,98

Tabel 6: de alternatieven van gewichten voorzien voor de NARCIS projectleiding

de formule van het gewicht van alternatief A_j is

$$\sum_{i=1}^{10} S'_{(A_j, C_i)} * w_{(C_i)}$$

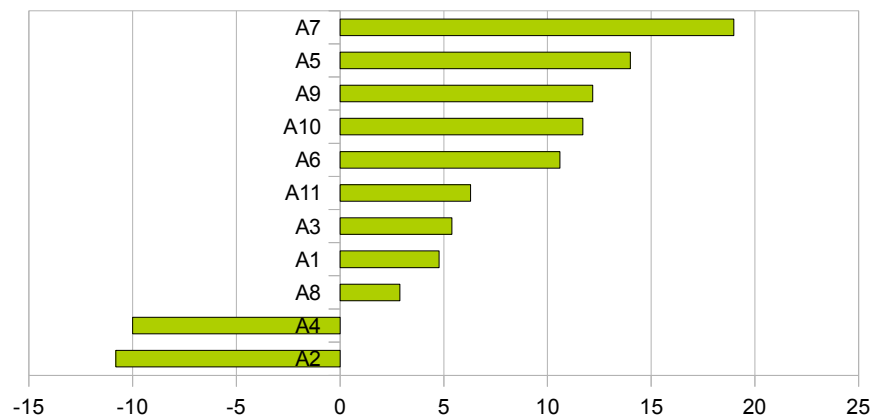
waarbij S' de genormaliseerde score S uit tabel 5 is, en w de weegfactor voor criterium C_i.

Op de kruising van alternatief 2 (A2) en criterium 1 (C1) staat 0,6. Dit is de uitkomst van

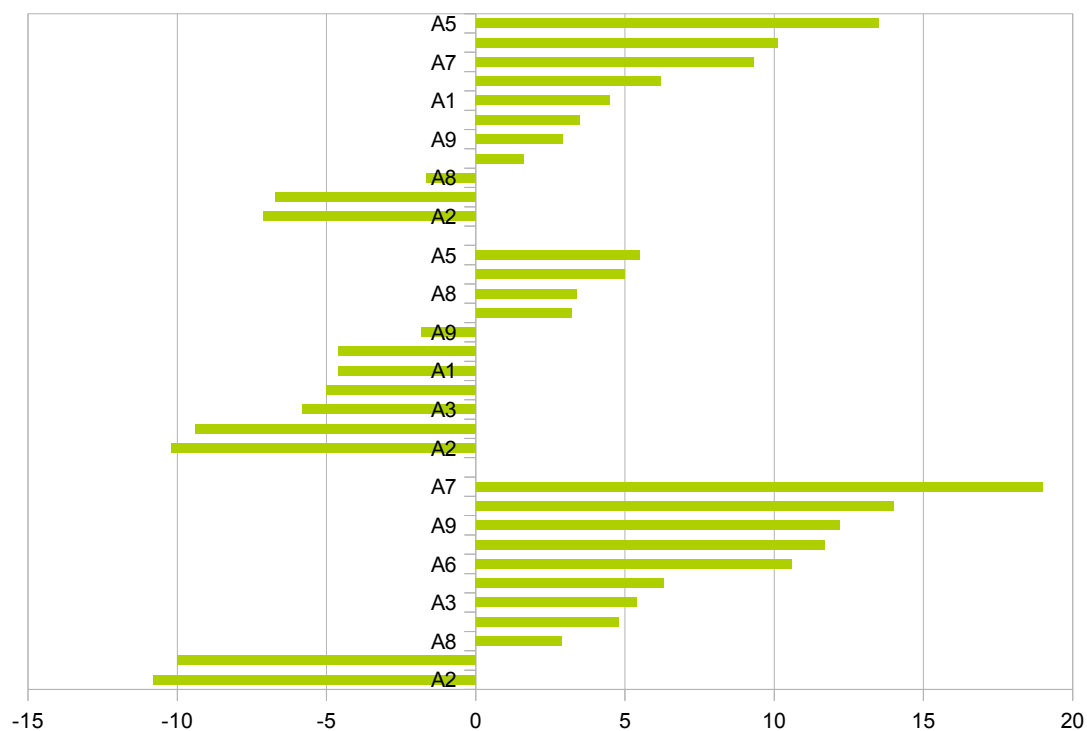
$$\left(\frac{10}{100} * 1\right) * 6$$

Hierbij is 10 de S uit tabel 5, dit was op een schaal van 1 tot 100, dus deze wordt genormaliseerd door te delen door 100 en vervolgens vermenigvuldigd met -1 of 1, afhankelijk of dit een positief of negatief criterium betreft. De 6 is w_(C1) uit tabel 4.

De sommatie over alle C levert gewicht -10,8 voor alternatief A2 voor de NARCIS projectleiding. Figuur 1 toont deze gewichten in een staafdiagram.



Figuur 1: zelfde waarden als tabel 6, maar dan in een staafdiagram



Figuur 2: de gewichten voor de stakeholders, achtereenvolgens de NARCIS projectleiding, de beleidsmakers en de repository managers

Nadere uitwerking alternatief A5 uit onderzoeksrapport

De meerwaarde van de Digital Author Identification

Coworkers

De omschrijving uit het onderzoeksrapport luidt:

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.

In dit rapport wordt dit nader uitgewerkt. De requirements worden bepaald en een mogelijke implementatie voorgesteld.

Requirements

1. [M] de visualisatie moet overzichtelijk zijn
2. [M] de visualisaties mogen onderling niet teveel verschillen qua look-and-feel
3. [M] de visualisatie moet passen in de NARCIS webomgeving
4. [M] het *kan* ingepast worden in de NARCIS webomgeving
5. [M] het visualisatie type moet schaalbaar zijn
6. [C] het genereren van de code mag niet te lang duren
7. [M] het moet duidelijk zijn wat er getoond wordt
8. [S] de onderlinge relaties tussen de auteurs is zichtbaar
9. [W] er wordt rekening gehouden met het tijdsaspect

Niet al deze requirements hebben dezelfde prioriteit. Daarom zijn ze gelabelled met de MoSKoW termen.

Data

De data uit de NOD met betrekking tot deelnemers aan een onderzoek is zelden compleet. Vaak wordt alleen de projectleider genoemd. Wanneer het een promotie-onderzoek betreft, worden bovendien meestal wel de promotors en de promovendus vermeld.

Voorlopig dus alleen afgaan op de publicaties.

Wat zijn de worstcase scenario's?

- grote groep
- een 'hechte' groep, waarin iedereen het met iedereen heeft gedaan.

- 1 (of meer) coworker die in meerdere subgroepen zit, niet in alle

Voorbeeld case

gegeven de volgende publicaties van Aalbers:

- pub1: Bladvlekkenziekte weer toenemend probleem door vochtiger klimaat (2010)
S1: {Aalbers, Deurloo, Mustert, Kloosterman}
- pub2: Pythium kan voor ernstige problemen zorgen (2007)
S2: {Aalbers, Kloosterman, Ostendorf}
- pub3: Opsporen oorzaak zwarte spikkels in radijs (2006)
S3: {Aalbers, Mustert, Caljouw}

De vereniging van de drie relaties S1, S2 en S3 noemen we S: {Aalbers, Deurloo, Mustert, Kloosterman, Ostendorf, Caljouw}

De PowerSet van S is de verzameling van alle subsets van S. S1, S2 en S3 zijn dus elementen van de PowerSet van S.

tekstueel

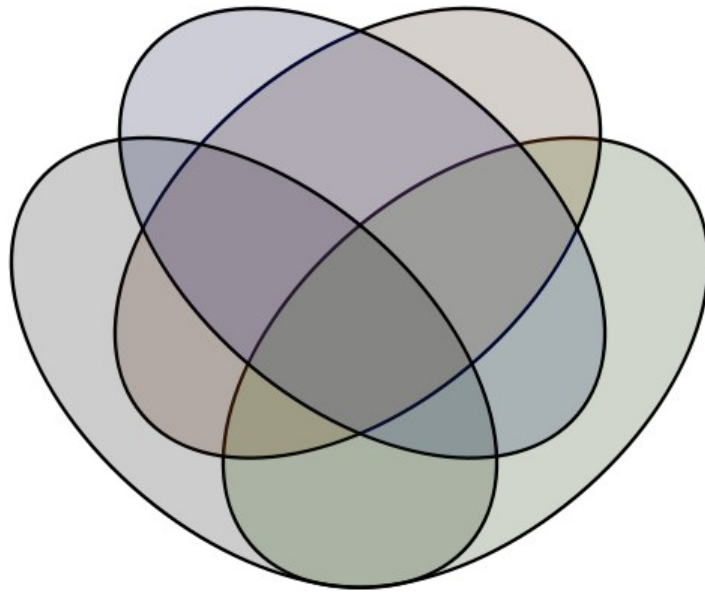
In tekst kan daar de volgende omgeving uit gedestilleerd worden:

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

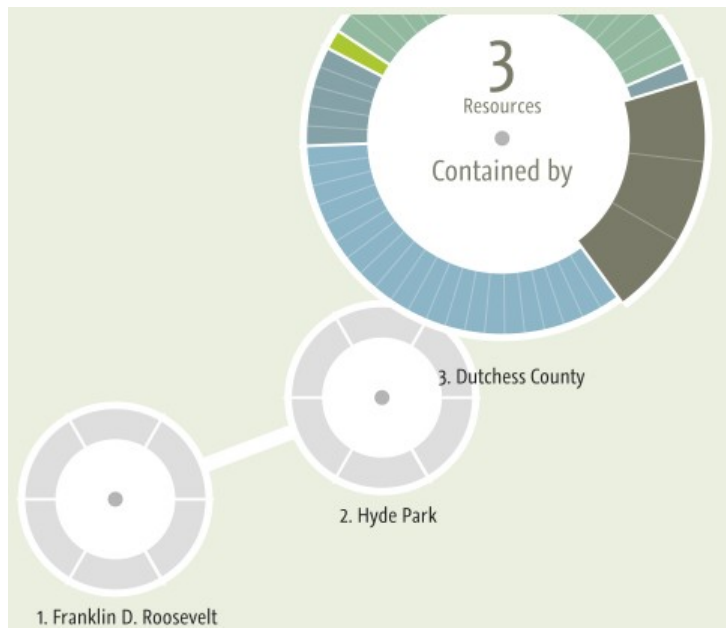
hierin zijn de onderlinge samenwerkingsverbanden verdwenen, alleen de relatie met de hoofdpersoon (Aalbers) bestaat nog.

De onderwerpen waarop de samenwerking heeft plaatsgevonden, evenals het temporele aspect, zijn niet weergegeven.

Venn diagram



Visual browser



Figuur 1: <http://askken.herokuapp.com/>

Als de persoon de eerste cirkel is, met de publicaties als punten op de cirkel, en de tweede cirkel is een gekozen publicatie, met de auteurs als punten op de cirkel is dit feitelijk een hypertree.

In deze representatie kun dezelfde auteur op meerdere plekken (vanuit meerdere punten uit de eerste cirkel) gelinked zijn. Door deze versnippering geeft het geen goed overzicht van de relatie tussen deze auteur en de hoofdpersoon.



Figuur 2: hyperbolic tree oftewel een hypertree

Een andere manier zou zijn om de eerste cirkel te laten bestaan uit de auteurs, en de tweede uit de publicaties van die auteur met de hoofdpersoon.

De onderlinge relaties tussen de auteurs zijn hier niet zichtbaar.

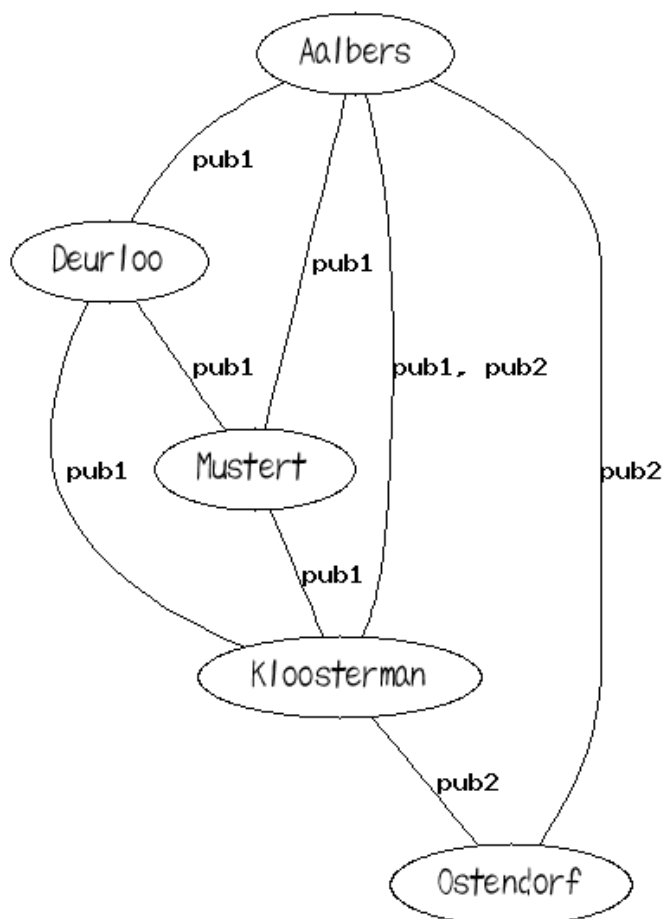
GraphViz

Een andere manier om deze data weer te geven is met behulp van een graph. Hierin zijn de nodes de personen, en de edges geven een auteur-relatie weer. S2: {Aalbers, Kloosterman, Ostendorf} wordt weergegeven door de relaties: {(Aalbers, Kloosterman), (Aalbers, Ostendorf), (Kloosterman, Oostendorf)}

de volgende dot-file (volgens GraphViz)

```
graph{
edge [dir=none]
Aalbers -> Deurloo [label=pub1]
Deurloo -> Mustert [label=pub1]
Deurloo -> Kloosterman [label=pub1]
Mustert -> Kloosterman [label=pub1]
Aalbers -> Mustert [label=pub1]
Aalbers -> Kloosterman [label="pub1, pub2"]
Aalbers -> Ostendorf[label=pub2]
Kloosterman -> Ostendorf [label=pub2]
}
```

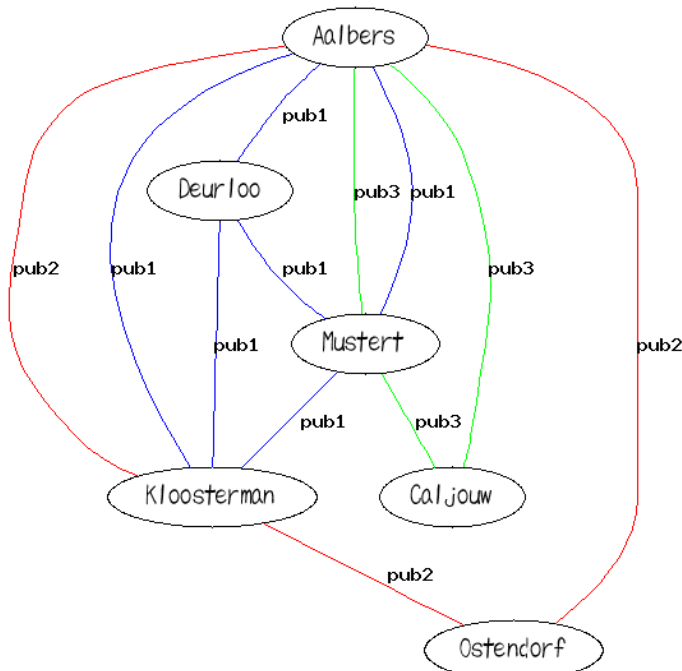
en het visuele equivalent:



Figuur 3: De relaties S1 en S2 weergegeven als graph

getekend door <http://ashitani.jp/gv/>

Om de clusters aan te geven, zou bijvoorbeeld kleur gebruikt kunnen worden. Dan wordt de relatie Aalbers-Kloosterman met het label 'pub1, pub2' opgesplitst in twee relaties Aalbers-Kloosterman met label 'pub1' en Aalbers-Kloosterman met label 'pub2'.

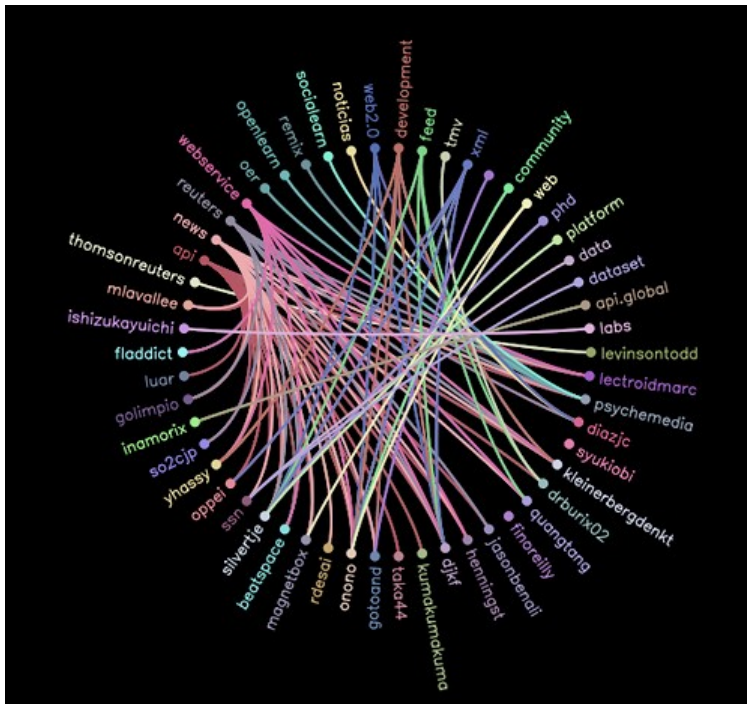


Figuur 4: De relaties S1 (blauw) , S2 (rood) en S3 (groen)

In simpele gevallen levert dit overzichtelijke graphs op, maar in de gecompliceerdere gevallen, wanneer het aantal coworkers groter wordt, worden de graphs onoverzichtelijk en beantwoorden ze niet aan hun doel, namelijk duidelijkheid verschaffen over de academische omgeving van een auteur.

MooWheel

Bovendien zijn de graphs onderling zeer verschillend. Een manier om dat te ondervangen is om de nodes in een vaste figuur te plaatsen, bijvoorbeeld een cirkel.

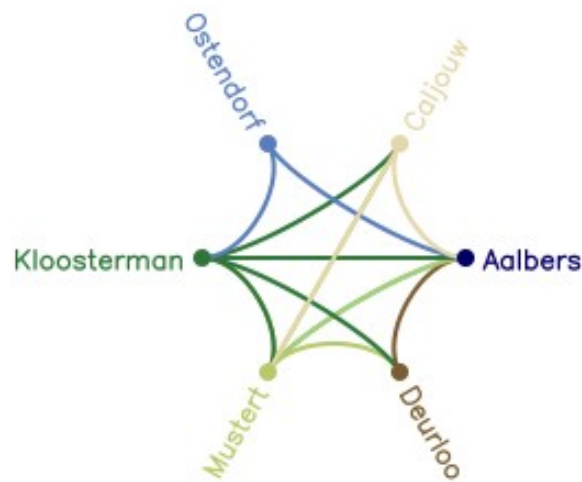


Figuur 5: moowheel gedownload van:
<http://ouseful.open.ac.uk/blogarchive/>



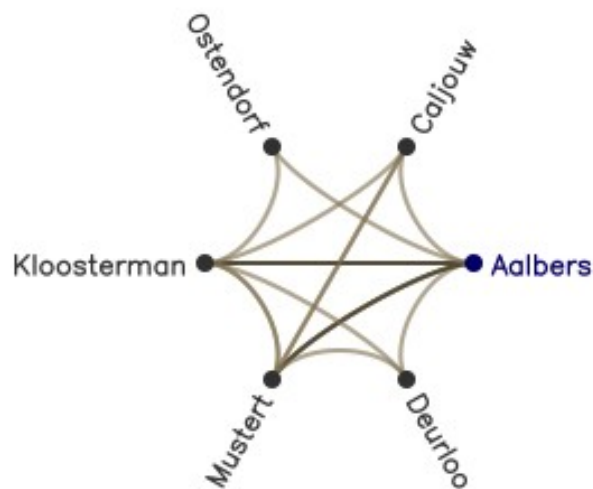
Figuur 6: gedownload van:
<http://www.datavisualization.ch/wp-content/uploads/2009/07/moowheel.png>

De look-and-feel is voor groepen met verschillende groottes min of meer gelijk. De S-relaties kunnen niet door middel van kleuren aangegeven worden: tussen twee nodes kan maar 1 edge getekend worden. De interne samenhang van de S-relaties wordt dus niet behouden.



Figuur 7: De relaties van Aalbers, weergegeven in een moowheel

Wel zou het aantal samenwerkingen dmv kleur weergegeven kunnen worden. Hoe meer gezamenlijke publicaties, hoe donkerder de edge. De kleuren komen dan voor de duidelijkheid uit 1 range.

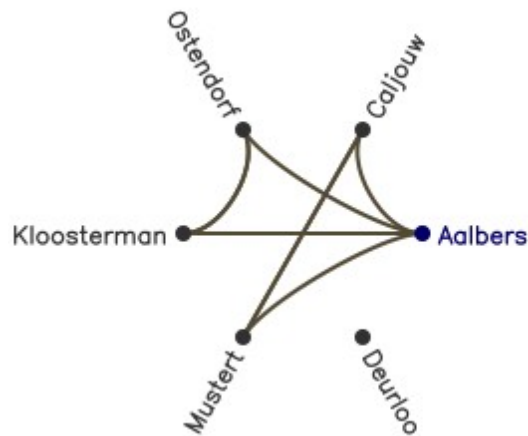


Figuur 8: moowheel van Aalbers waarbij de kleur een indicatie is voor het aantal gezamenlijke publicaties

Het temporele aspect kan tot uitdrukking gebracht worden door overlays van de verschillende jaren te tonen, wellicht met een soort onion skinning¹. Dit zou gedaan kunnen worden in combinatie met 4.2.6, de academische loopbaan van een auteur.

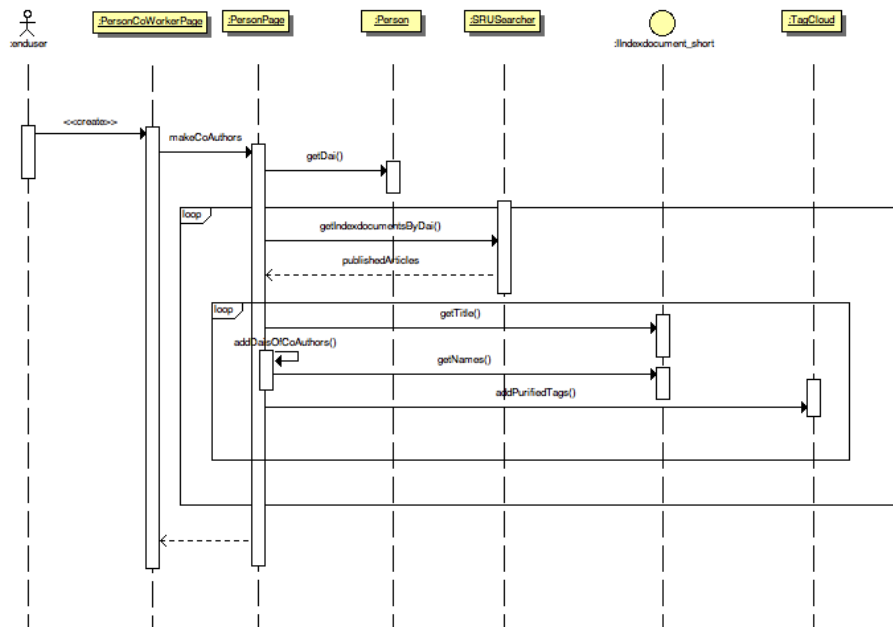
¹ Een term uit de 2D computer graphics, waarbij tijdens het maken van een animatie de voorgaande frames 'erdoorheen' schijnen. Ook gebruikt bij stop motion.

Wanneer een jaar geselecteerd is, kan het wheel t/m dat jaar getekend worden. De toekomstige auteur-nodes kunnen al wel getekend worden, maar dan zonder edges.



Figuur 9: de samenwerkingen van Aalbers tot 2007

Het maken van de coauthors bij een PersonPage



Figuur 10: sequence diagram van de makeCoAuthor methode

afgeleide kennis

idee

visualiseer een netwerk
van een onderzoeker

$$S : \{ p \mid p.dai \vee p.PRS \}$$

de verzameling S van alle personen uit narcis
die een dai of een PRS hebben

$$\wp(S)$$

de powerset van S

$$R(S) \subseteq \wp(S)$$

de relatie R tussen p's uit S

waarvoor geldt dat er een relatie beschreven is
dmv een *publicatie* of een *onderzoek**

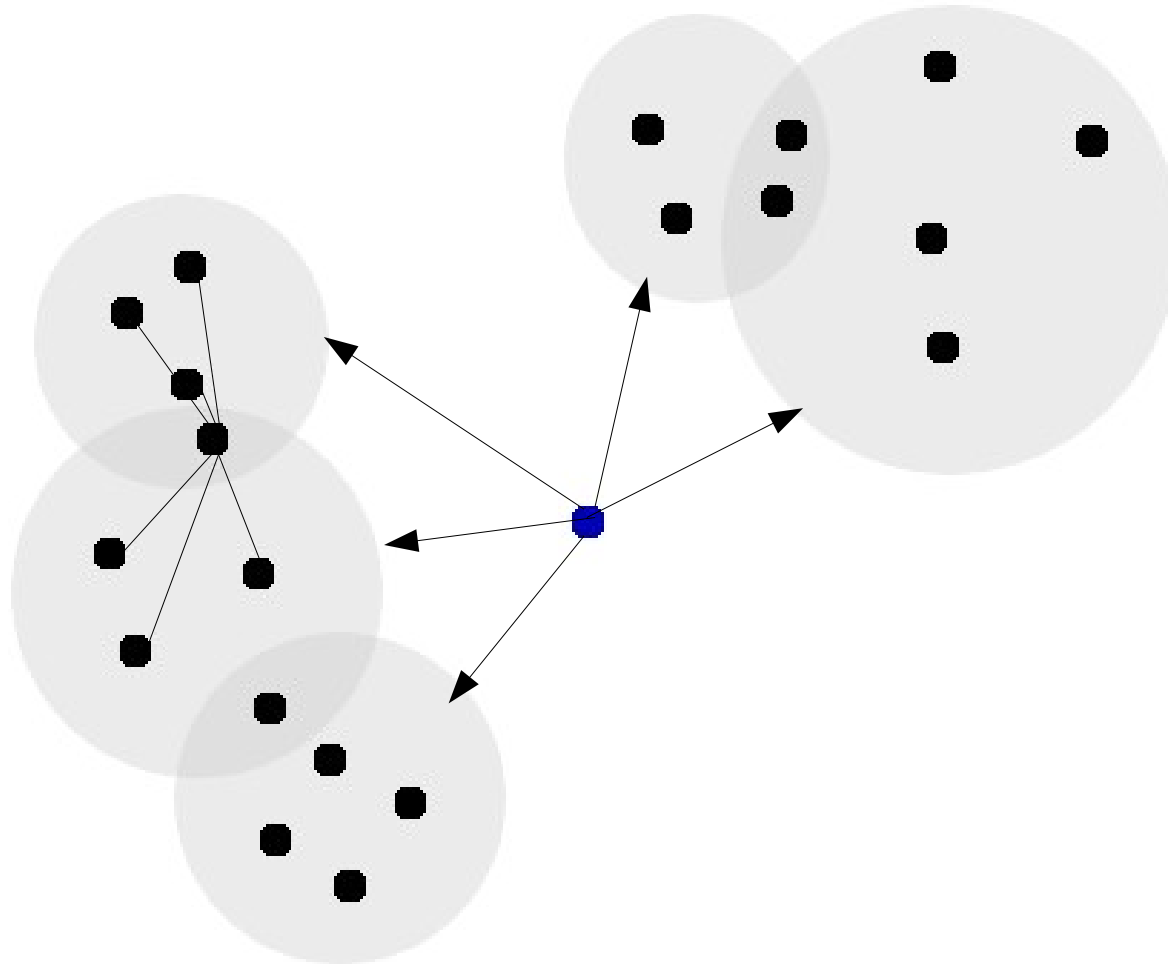
$$R_{bennis}(S): \left\{ r \in R(S) \mid p_{bennis} \in r \right\}$$

voor persoon Bennis:

de verzameling van alle relaties uit de
bestaande relaties waar Bennis een onderdeel
van uitmaakt

Graph Theory

- nodes
- edges



idealiter

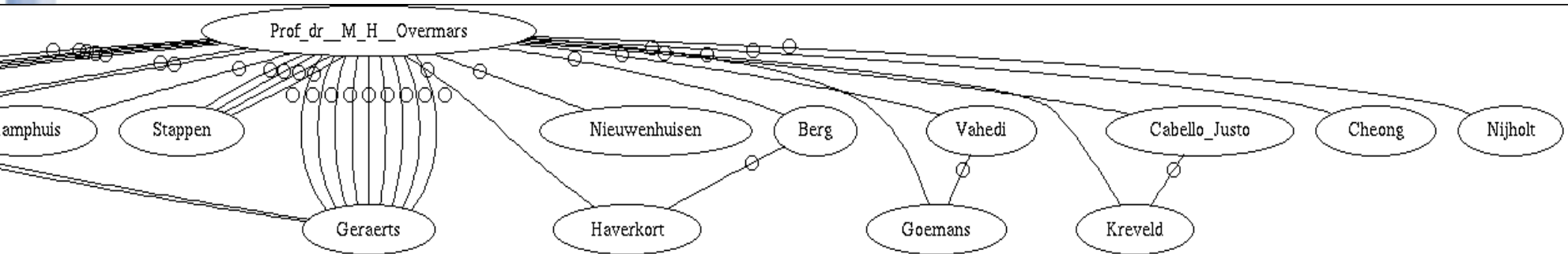
- wat als p in meer dan 1 overlap zit?
- multi-Venn diagrammen zijn mogelijk, maar verliezen dan hun eenvoud en dus gebruiksgemak voor overzicht



wensenlijst

- schaalbaar
- overzichtelijk
- 'uniforme' layout
- clustering
- 'eerlijk'
- gewicht van relaties

GraphViz



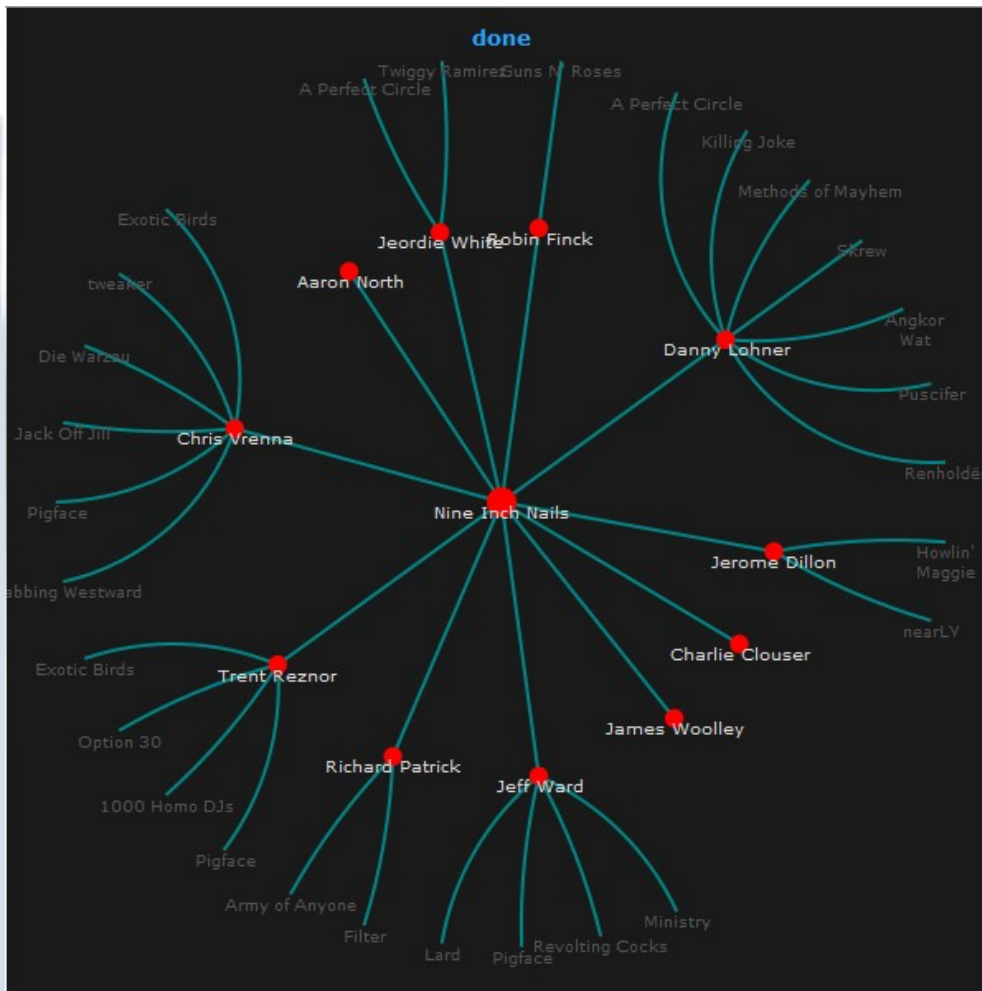
- voordeel

- flexibel
- multi-purpose
- recht toe, recht aan

- nadeel

- slecht schaalbaar
- tekenalgoritme

hypertree



- voordeel
 - schaalbaar
 - clustering (waar mogelijk)
 - proximity
- nadeel
 - boomstructuur (directed graph)
→ geen overlap

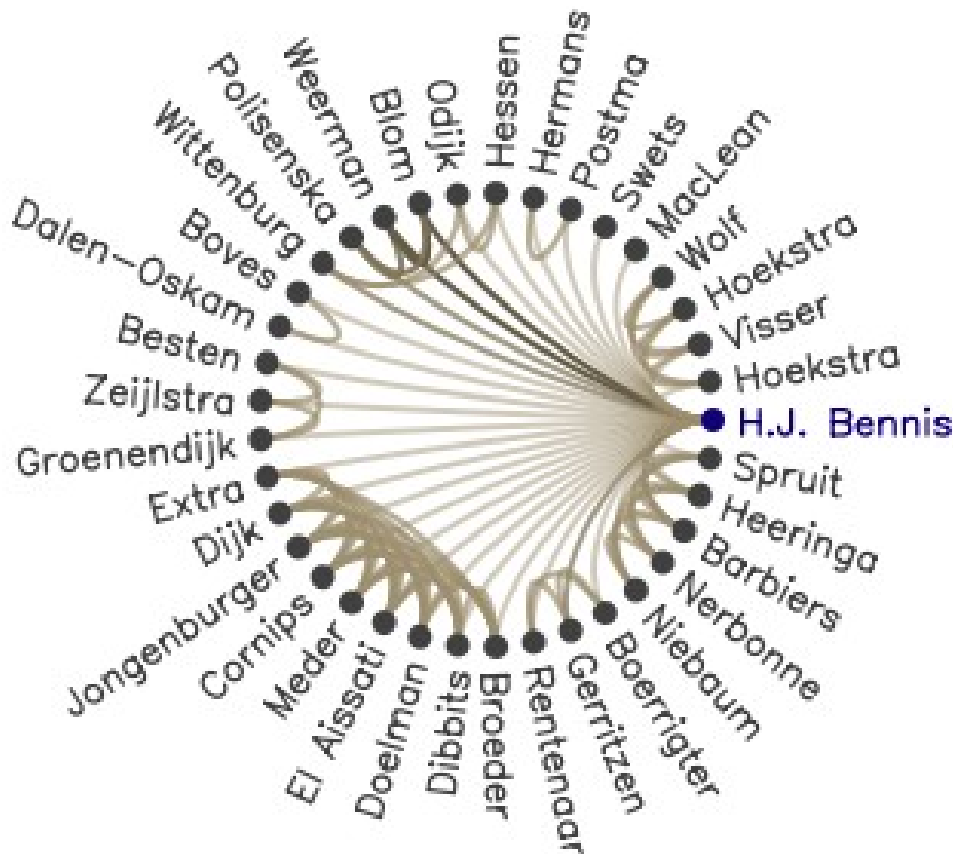
outerplanar graph

- voordeel

- schaalbaar
- clustering

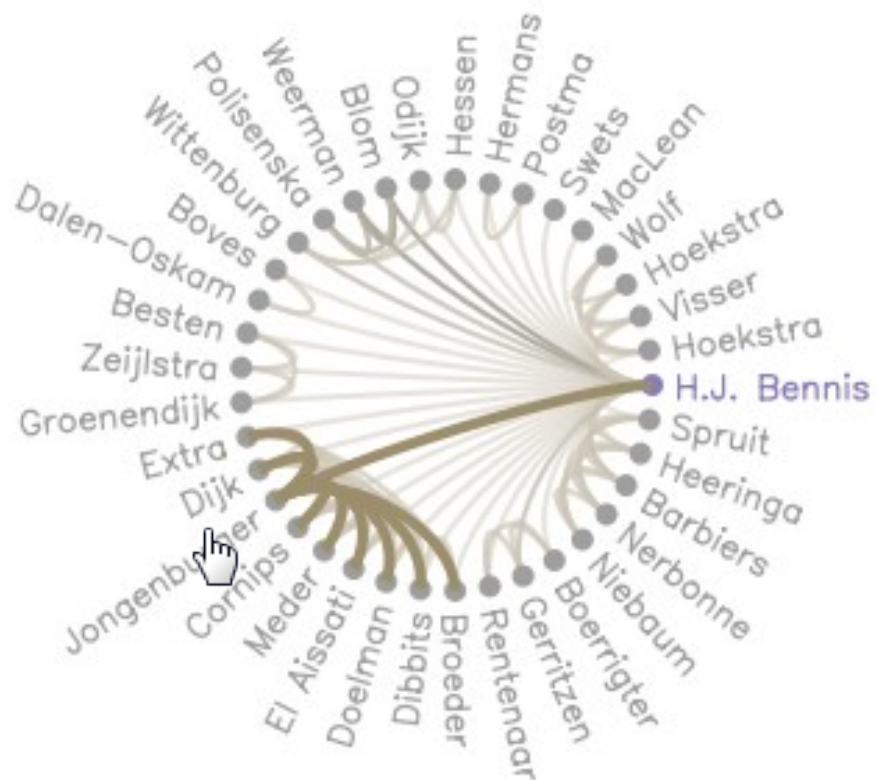
- nadeel

- bij veel edges crowded
- proximity betekent niet altijd hetzelfde



verband cluster ↔ keywords

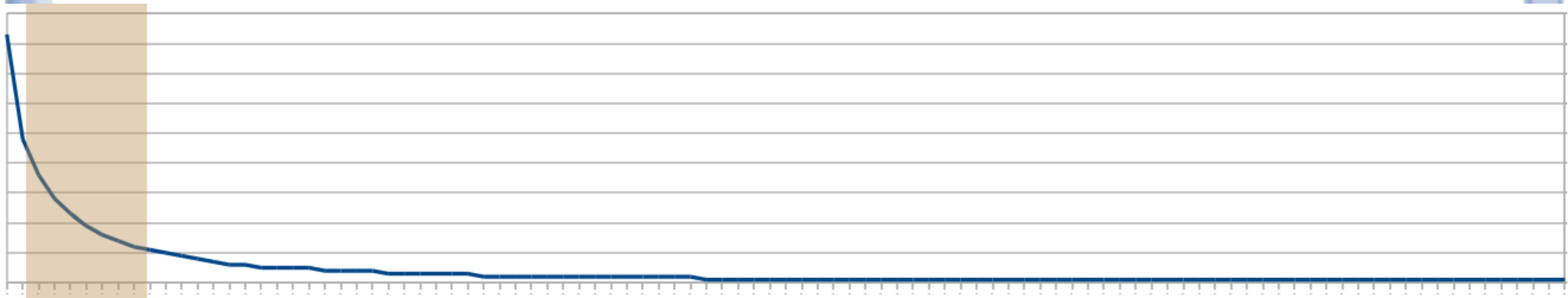
CHILDREN CITY CULTURES GOVERNMENT INFRASTRUCTURE INTERFACE LINGUISTIC
NEIGHBOURHOOD RESOURCES SAND **SYNTAX** TEXTS URBAN



woordenwolk

- welke woorden zijn interessant:
 - niet te gewoon:
 - these, which, studies, analysis
 - zijn niet beschrijvend genoeg
 - niet te zeldzaam:
 - extralegal, threonine, Hsul, uncoating
 - zijn te specifiek

woordenwolk



het aantal keer dat een woord
voorkomt in de populatie





zo erg is het gelukkig niet

Nadere uitwerking alternatief A7 uit onderzoeksrapport de meerwaarde van de Digital Author Identification

Verrijkte Publicaties

De omschrijving uit het onderzoeksrapport luidt:

“Het idee is om onderzoeken uit de NOD te koppelen aan publicaties en datasets. Dit is een lastige bezigheid, er moet gematched worden op titel, samenvatting, auteur, etc. Bijvoorbeeld: onderzoekers die gekoppeld zijn aan een onderzoek in de NOD, en samen een publicatie hebben ten tijde van dat onderzoek. Dit kan een mismatch opleveren.

Een manier om de precision [...] te verhogen is het verkleinen van de pool van publicaties tot alleen doctoral theses.

Promotieonderzoek heeft vaak dezelfde titel als de bijbehorende thesis. Om vast te stellen of een specifiek onderzoek een promotie-onderzoek betreft, kan gekeken worden naar het attribuut isThesis. Bij de NOD-tabel werkzaamheid wordt bijgehouden in welke rol een persoon bij een onderzoek betrokken is. Wanneer dit als promovendus of promotor is, kan worden aangenomen dat het onderzoek een promotie-onderzoek is. Dit geldt voor zo'n vijftienduizend onderzoeken.

Wanneer de DAI van de promovendus gelijk is aan de DAI van de auteur van het proefschrift, kan aangenomen worden dat die bij elkaar horen.” [TJR01]

Citaat 1: Demonstrator II: promotie-onderzoek

In dit rapport wordt dit nader uitgewerkt. De requirements worden bepaald en een mogelijke implementatie voorgesteld.

Requirements

1. [M] de visualisatie moet overzichtelijk zijn
2. [M] de gelegde relaties moeten *waar* zijn
3. [M] het *kan* ingepast worden in de NARCIS webomgeving
4. [C] de bekende data *moet* zoveel mogelijk gemapped worden op de ontologien
 - 4.1. [M] onderzoek en daaruit voortgekomen publicaties staan centraal
 - 4.2. [M] alleen personen met een DAI worden bekeken
 - 4.2.1. [M] een auteur is of gelabelled met dc:creator, of heeft een auteurs-rol in deze deelset van MARCcodes: {aut, edt, dis, pta, rev, ctb, cre, prg}
 - 4.2.2. [C] wanneer een onderzoeker geen DAI heeft, maar wel een PRS kan die identificatie gebruikt worden
 - 4.3. [C] de uitvoerende organisatie bij een onderzoek moet getoond worden
 - 4.3.1. [W] de uitvoerende organisatie wordt NIET gematched op de organisatie uit de publicaties (mochten die genoemd worden)
5. [M] de relatie tussen een onderzoek, de onderzoeker en de bijbehorende publicatie moet zichtbaar zijn
6. [M] uitgaande van onderzoek en publicatie is elke relatie die met 1 stap te benaderen is, ook present in de RDF, voor zover hier een ontologische representant voor bestaat.
 - 6.1. [M] Bij personen *moet* tenminste de naam, de functie binnen de aggregatie en de werkbetrekking getoond worden.

Niet al deze requirements hebben dezelfde prioriteit. Daarom zijn ze gelabelled met de MoSKoW termen.

Data

De data uit de NOD met betrekking tot deelnemers aan een onderzoek is zelden compleet. Wanneer het een promotie-onderzoek betreft, worden in ieder geval wel de promotors en de promovendus vermeld. Bovendien is de titel bekend.

Bij doctoralThesis publicaties worden de promotors in MODS met een MARC-code aangegeven. Deze promotors zijn vaak voorzien van een DAI. De promovendus zelf vaak niet, dit is nog aan het begin van de wetenschappelijke carrière.

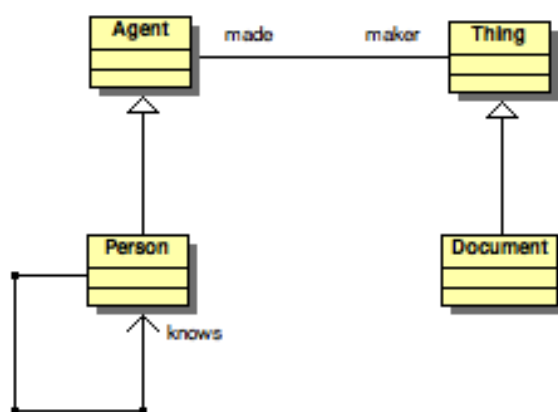
Dit is dus een subset uit de onderzoeken en publicaties die gemakkelijker gematched kan worden, aangezien de kans groter is dat de benodigde data aanwezig is.

Een van de toegestane ontologiën is Friend Of A Friend (FOAF)¹. Hierin kunnen relaties tussen personen onderling en tussen personen enerzijds en projecten, organisaties en documenten anderzijds gelegd worden.

¹ <http://xmlns.com/foaf/spec/>

Hoewel hier dezelfde typering lijkt te zijn als binnen NARCIS, kunnen niet alle relaties gelegd worden die binnen NARCIS bestaan. Bijvoorbeeld de relatie tussen een Person en een Document is binnen NARCIS uitgebreid beschreven door middel van de MARC Code voor Relators². In foaf is de enige relatie die gelegd kan worden de maker-relatie, zoals getoond in figuur 1.

De Universiteit Twente heeft een aantal ontologien ontwikkeld die beter aansluiten bij de mods-praktijk. Hiermee kunnen documenten getypeerd worden, kunnen relaties gelegd worden tussen topics en personen



Figuur 1: foaf-relatie tussen een Person en een Document

² <http://id.loc.gov/vocabulary/relators.html>