Methodologische vernieuwing en bronnenkritiek in het digitale tijdperk - 2018-2023

Rik Hoekstra, Marijn Koolen, Sebastiaan Derks, Jan Burgers, Marijke van Faassen en Ida Nijenhuis

Inleiding: dataclouds en methodologische uitdagingen

Enorme hoeveelheden informatie over het verleden komen beschikbaar in digitale vorm: zoveel dat het vrijwel onmogelijk is om deze in samenhang te onderzoeken zonder de inzet van digitale technieken. Deze digitale wending brengt fundamentele transformaties teweeg in de geesteswetenschappelijke onderzoekspraktijk, waaraan tot op heden nog te weinig aandacht is besteed.

Digitale onderzoeksgegevens worden in toenemende mate samengebracht binnen gedeelde data-infrastructuren. Dergelijke datawolken, zelfs van een beperkte omvang, dragen echter het gevaar in zich dat ze een amorf en ondoordringbaar karakter krijgen, die buitengewoon lastig te bevragen en te doorgronden zijn. Dit stelt geesteswetenschappelijke onderzoekers voor belangrijke methodologische en onderzoekspraktische kwesties.

Zo is de context van digitale bronnen en gegevens in veel gevallen anders en veel complexer dan bij analoge bronnen. Hoe kunnen we daar het beste mee omgaan en de contextuele informatie zoveel mogelijk waarborgen? Hoe zorgen we ervoor dat we zicht houden op de ontstaansgeschiedenis, representativiteit en structuur van de bronnen? Hoe voorkomen we dat de inzet van digitale tools tot steeds groeiende black boxes leidt? Door data-linking kunnen veel gegevens met elkaar worden verbonden, maar wat is de inhoudelijke betekenis van deze koppelingen, en wat zijn de best practices hierbij? Hoe kunnen we ons arsenaal aan methoden uitbreiden en deze overdragen op nieuwe generaties, zodat het huidige onderzoek op de lange termijn toegankelijk, herhaalbaar en begrijpelijk blijft?

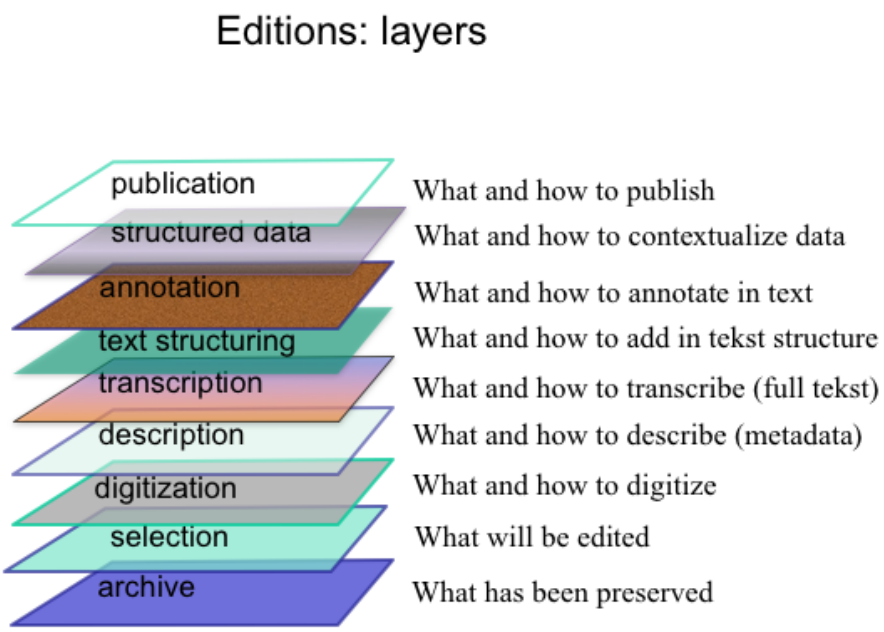
Dit zijn enkele van de methodologische vragen waarvoor onderzoekers zich gesteld zien; zij kunnen deze vragen vaak niet alleen adresseren. Gezien de complexiteit van de methodologische vragen en de voortdurende evolutie van de digitale infrastructuur, zijn blauwdrukken voor deze noodzakelijke methodologische vernieuwing moeilijk te geven – het is een doorlopende ontwikkeling waarvoor het Huygens ING wel goed geëquipeerd is, maar waarvoor nog geen heldere oplossingsroutes kunnen worden aangewezen We pretenderen dan ook niet de oplossing te hebben voor deze zeer ingewikkelde en veelomvattende problematiek, die bovendien veel breder is dan de vakgebieden waarmee het Huygens ING of zelfs het Humanities Cluster zich bezig houdt. Wel kunnen we een aantal methodologische handvatten geven die voortkomen uit onze ervaringen en expertise op het terrein van digitale bronnen, teksten en data. Daarmee kan een begin worden gemaakt met een aanpak die zich hoofdzakelijk richt op de specifieke problemen van de onderzoeksvelden van het instituut; het is voorts duidelijk dat deze interdisciplinaire inbreng behoeft, waarbij zowel het humanitiesonderzoek als dataspecialisten en informatie-technologen zijn betrokken.

Methodologische aspecten

1. Vernieuwing van de bronnenkritiek

Sinds het ontstaan van de historisch gerichte wetenschappen geldt dat onderzoekers weinig met hun materiaal kunnen beginnen zonder bronnenkritiek. De vaardigheid om bronnen op hun waarde en waarachtigheid te beoordelen is immers een essentieel kenmerk van wetenschappelijk onderzoek. De complexiteit van de digitale bronnenverzamelingen maakt kritisch bronnenonderzoek steeds complexer en vereist nieuwe vaardigheden.
Alleen de meest triviale vragen kunnen namelijk direct beantwoord worden uit de beschikbare data. Met de huidige stand van zaken van techniek en digitalisering is het weliswaar eenvoudig geworden de spreekwoordelijke speld in de hooiberg te vinden, maar met wetenschappelijke vragen die voortkomen uit verwondering, die het waard zijn om te dienen als kernvragen voor een onderzoeksproject, is dit zelden het geval. Om die uit gegevens te beantwoorden, moeten ze vertaald worden naar deelvragen die elk weer met eigen aanpak en technieken moeten worden beantwoord.

Daartoe is het zinnig samengestelde (digitale) bronnen en data op formele gronden te ontleden in de verschillende aspecten die aan hun totstandkoming hebben bijgedragen. Als analytisch model stellen we een lagenmodel voor uitgegeven bronnen voor, dat kan dienen om de verschillende aspecten te onderscheiden. Omdat het een analytisch hulpmiddel is, is het niet altijd makkelijk de werkelijkheid er precies mee te omschrijven: in de praktijk zijn de grenzen tussen verschillende lagen niet altijd scherp. Ook zijn ze bij lang niet alle bronnenuitgaven of datasets aanwezig: soms zijn er, om een simpel voorbeeld te noemen, alleen digitale facsimile voorhanden en geen transcripties, andere keren zijn er wel transcripties maar ontbreken afbeeldingen of betrefwoording. In een bron kan iedere willekeurige samenstelling van lagen voorkomen. Grosso modo geeft dit lagenmodel wel de volgtijdelijkheid van de bewerkingen aan, en soms is er ook een afhankelijkheid. De selectie en de publicatie zijn de enige lagen die altijd in een of andere vorm aanwezig zijn.



2. Combinatie van gegevens

De onderzoeker moet steeds bronnen en data met elkaar combineren en vaak op een nieuwe manier. Dat stuit op nieuwe moeilijkheden, want zelfs als verschillende bronnen of datasets aanvullende gegevens bevatten

over dezelfde fenomenen, zoals personen, plaatsen of gebeurtenissen, dan is dat vaak op verschillende manieren, met verschillende interpretaties, aanduidingen en verwijzingen. Er zijn dan steeds bewerkingsslagen nodig voordat nuttige koppelingen gemaakt kunnen worden.
De contextualisering en verrijking van bronnen vindt in toenemende mate plaats in interactie met de bestaande digitale infrastructuur en heeft voor ontsluiting andere consequenties dan voor analyse.

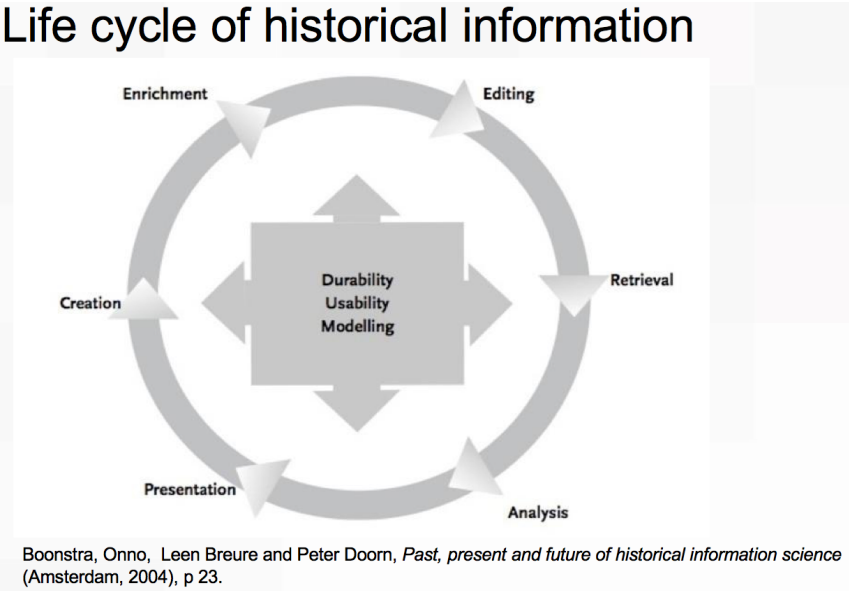
- Ontsluiting binnen een infrastructuur betekent veel meer dan voorheen dat al of niet conformeren aan bestaande afspraken consequenties heeft voor de toegankelijkheid van bronnen. Optimale ontsluiting (en daarmee automatisch contextualisering) vereist dat binnen de infrastructuur conventies in structuur en inhoud worden gehanteerd. Voor adoptie daarvan moeten deze 'enabling' zijn en geen keurslijf. Complexen van samenhangende gegevens krijgen door de combinatie van talrijke facetten kaleidoskopische herkomstinformatie en schakeringen die snel ondoorzichtig kunnen worden. De technische organisatie en structurering van de gegevens moet het mogelijk maken de gegevens traceerbaar te houden, maar onderzoekers en dataspecialisten moeten ze zodanig representeren en onderbrengen dat ze ook echt bij elkaar aansluiten.
- Voor analytisch onderzoek vereist werken binnen een infrastructuur dat onderzoekers zich bewust zijn van de mogelijkheden, beperkingen en eigenaardigheden van de context waarbinnen hun onderzoeksmateriaal zich bevindt. Dat vereist een aantal vaardigheden waaronder a) de kritische beschouwing van de samenstelling van de aanwezige data en hun herkomst en totstandkoming, b) de vaardigheid de beschikbare informatie optimaal te bevragen en te verwerken, en c) inzicht in hoe aanvullende, externe data zodanig te combineren is dat zowel de context als de toegevoegde data recht worden gedaan.

In beide gevallen, ontsluiting en analyse, kan het lagenmodel als analytisch hulpmiddel goede diensten bewijzen.

3. Transparant maken van het iteratieve onderzoeksproces

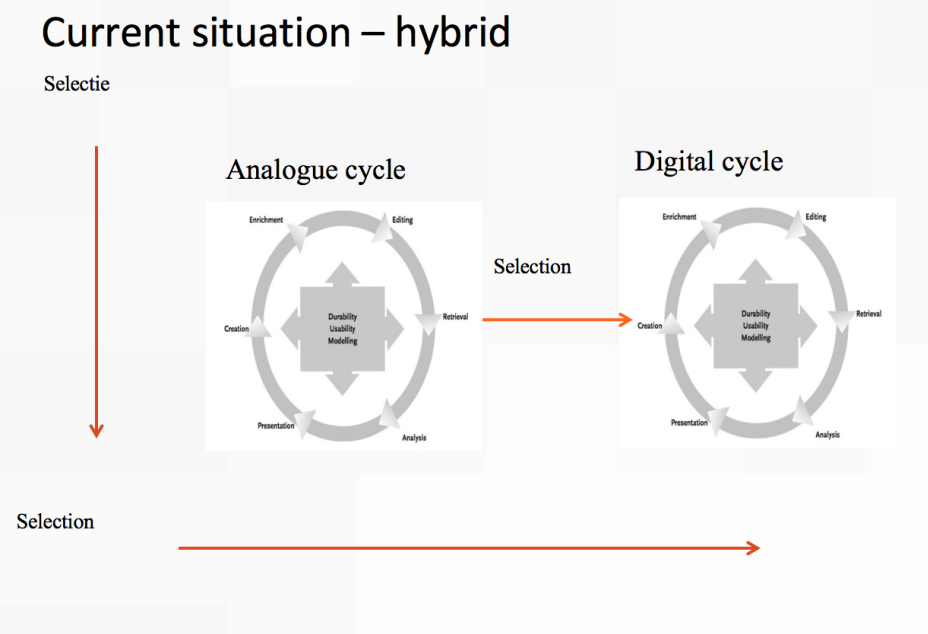
Resultaten die verkregen worden uit zoekopdrachten in digitale bestanden, roepen nieuwe vragen op, naar context of naar representativiteit. Vaak blijkt dat een eerste onderzoeksvraag helemaal niet beantwoord kan worden, omdat het antwoord niet direct uit de beschikbare bronnen af te leiden is. Als de onderzoeker geluk heeft, wordt dit probleem snel duidelijk, maar voor een meer adequate beantwoording is wel kennis nodig over hoe de gebruikte gegevens in elkaar zitten en wat hun context is. En die kennis is alleen te verkrijgen of over te dragen door directe omgang met de gegevens en inzicht in hun ontstaansgeschiedenis en structuur.

Voor een beschrijving van het proces van informatieverwerking wordt vaak het 'life cycle of historical information' model van Doorn, Breure en Boonstra gebruikt, dat hieronder is weergegeven.



Boonstra, Onno, Leen Breure and Peter Doorn, *Past, present and future of historical information science* (Amsterdam, 2004), p 23.

Wij stellen voor het model voor de omgang met complexe data uit te breiden voor de huidige, hybride situatie waarin onderzoekers zowel analoge, als gedigitaliseerde, als *born digital* bronnen en data gebruiken in hun onderzoek. In deze hybride situatie is sprake van een verdubbeling van de cirkel, die vooral wordt gecompliceerd door verdubbeling van de selectiemechanismen en ook wel door de wijze waarop het originele materiaal vaak dubbel wordt bewerkt. Als bijvoorbeeld een bestaande, gedrukte bron wordt omgezet voor gebruik in een digitale omgeving dan geldt nog steeds dat de selectie- en bewerkingsmethoden van de originele bewerkers hun stempel hebben achtergelaten op de data, maar dat (vaak) een deel van de gedrukte context verloren gaat, al was het maar omdat de fysieke indeling (en inhoudsopgave) verandert. Daar komt bij dat de wijze van digitaliseren en bewerken er extra lagen aan toevoegen die voor de onderzoeker kunnen dienen als hulpmiddel voor de toegankelijkheid, maar ook het zicht op het origineel verder kunnen vertroebelen.



Het spreekt voor zich, dat dit probleem groter wordt naarmate er meer data worden gecombineerd: enerzijds groeit de context en daarmee worden de data rijker, maar anderzijds wordt de context opgebouwd uit aspecten met een steeds heterogenere achtergrond. Zowel de bewerking als het gebruik van bronnen en data wordt hiermee een iteratief proces, dat de data verrijkt, maar ook verandert.

In deze situatie wordt het steeds belangrijker dat zowel bewerkers van data als onderzoekers die werken met die data, hun werkwijze transparant houden en daarvan verslag leggen, zodat niet alleen de resultaten overdraagbaar en controleerbaar blijven, maar ook dat zij zelf het proces kunnen bijhouden. Dit vraagt om een samenhangende methodologie, dat wij willen vangen onder het concept van *data scopes*.

Data scopes

Vaak zijn er vele databewerkingen nodig om de data geschikt te maken voor een onderzoeksvraag: bewerkingen om data op te schonen, te selecteren, te koppelen, en te presenteren. Al deze bewerkingen veranderen de data en vergen interpretatie van de data om een geschikte bewerking te kiezen, en zijn dus onderdeel van het onderzoek. Een concreet voorstel voor methodologische vernieuwing is het uitwerken van het concept *data scopes* als een instrument om dit databewerkingsproces transparant en herbruikbaar te maken, en aan te laten sluiten bij de scope van de onderzoeksvraag. Hiermee krijgen interdisciplinaire teams een handvat om, waar nodig, iteratief de onderzoeksscope en data scope op elkaar af te stemmen. De bredere onderzoeksgemeenschap kan aan de hand van een *data scope* beoordelen in welke mate het bewerkingsproces binnen een onderzoek aansluit bij de scope van de onderzoeksvraag. Daarnaast bieden *data scopes* een middel om te bespreken hoe databewerkingsprocessen opgenomen kunnen worden in nieuwe methoden die bruikbaar zijn voor de relevante onderzoeksvelden binnen de geesteswetenschappen.

Praktische invulling

Het Huygens ING is uitstekend toegerust voor deze vernieuwing van methodologie en bronnenkritiek. De hier verenigde deskundigheid maakt het mogelijk het onderzoeksproces te overzien, kritisch te onderzoeken en best practices te ontwikkelen. In eerste instantie kan dat het best door dit toe te passen op enkele lopende (kern)projecten van het instituut; zie voor enkele uitwerkingen bijlage A.

In de context van onderzoeksprojecten met een infrastructurele component hebben de volgende instrumenten hun waarde bewezen voor communicatie en concrete aanpak van onderzoeksvragen en aansluiting bij de infrastructuur:

- Use cases; alle betrokkenen werken uit hun eigen expertise vanuit zo nauw mogelijk omschreven gebruiksmogelijkheden van data, software of toepassingen. Deze methode heeft zich al bewezen als een goede manier om concreet en gelaagd informatie uit te wisselen tussen mensen met een andere achtergrond. Op deze manier kan onder meer het vraagstuk van de modellering worden aangepakt.
- Modelleren; dit is het maken van een representatie en abstractie van de 'echte wereld' in een vorm die geschikt is voor analyse, een cruciale functie in het onderzoek. Ook hierbij is veelvuldig overleg over onderzoeksvragen, (technische) beperkingen en mogelijkheden van groot belang. Modellen zouden expliciet benoemd en getest moeten worden.
- Tooling; alle betrokkenen maken gebruik van tools voor het bewerken, bevragen of analyseren van onderzoeksmateriaal. Alle tools hebben echter eigen karakteristieken en eigenaardigheden die ze geschikt maken voor de ene taak, maar minder geschikt voor een andere. Naast ICT-specialisten moeten ook dataspecialisten en onderzoekers een rol krijgen in toolkritiek. Deze toolkritiek moet worden geïntegreerd in bronnenkritiek.

Er zijn nieuwe samenwerkingsvormen nodig om dit aan te pakken. Door op bovengestelde manier samen te werken, en waar dat zinvol en mogelijk is ook met externe partners, wordt de methodologische vernieuwing gerealiseerd.

Beoogde resultaten

Hierboven zijn enkele methodologische uitdagingen aangegeven en richtingen voor werkvormen om deze nader te exploreren en naar oplossingen te zoeken om bestaande methoden te hergebruiken en aan te passen, maar ook uit te breiden met nieuwe methoden, die vaak tenminste deels zijn ontleend aan of rusten op computertechnieken en computer science.

Om de expertise te bundelen en de communicatie intern in het Huygens ING maar ook naar de andere instituten van het HuC en de buitenwereld te verbeteren stellen we voor de methodologische inspanningen in concrete producten vorm te geven:

- een website waarop best practices en methodologische discussies worden gepubliceerd. Op den duur kan dit gaan dienen als online handboek;
- training/onderwijs middels workshops over de drie thema's (1) vernieuwing bronnenkritiek, (2) combineren van gegevens, en (3) transparant maken van het onderzoeksproces;
- rapportage door (leden van) projectteams van bevindingen en ervaringen rondom de thema's, die vast onderdeel zouden moeten uitmaken van de wetenschappelijke output en projectrapportages

Bijlage A Concretisering aan de hand van enkele projecten

Projecten

Charterbank

In de charterbank worden in principe alle Nederlandse oorkonden opgenomen. Daartoe worden beschrijvingen geharvest uit allerhande archieven en waar mogelijk aangevuld met afbeeldingen van de charters. De beschrijvingen zijn heel beperkt omdat van de meeste maar heel weinig gegevens beschikbaar zijn in de oorspronkelijke archiefinventarissen. Door de combinatie van de charters uit allerlei verschillende vindplaatsen wordt dit al een unieke verzameling. Deze kan worden verrijkt door personen en plaatsnamen die over het algemeen in beschrijving worden genoemd maar niet systematisch zijn ontsloten, daaruit te verzamelen en ze te disambigueren. Uit de (vele) oorkondenboeken die het Huygens ING al digitaal beschikbaar heeft kunnen voorts de daarin opgenomen ontsluitingen en de transcriptie van de oorkonden worden toegevoegd om ze althans deels ook tekstueel toegankelijk te maken. Dit geldt ook voor de transcripties van oorkonden in de digitale Registers van de Hollandse grafelijkheid 1299-1345.
De verschillende disciplines moeten hier samenwerken om de namen te harvesten en te ontdubbelen. De toegankelijkheid van de charters wordt voorts verbeterd door de beschikbaarheid van zowel de persoonsnamen als de digitale tekst uit de oorkondenboeken, maar er moet duidelijk blijven dat dit maar voor een deel van het materiaal geldt en dat er grote onzekerheden zijn bij de interpretatie van veel namen. Hier moeten vooral dataspecialisten en onderzoekers samenwerken om dit ook aan de gebruikers duidelijk te maken.
Trefwoorden: harvesten, disambigueren

Resoluties Staten-Generaal

In het project Resoluties Staten-Generaal worden niet slechts de besluiten zelf gedigitaliseerd, maar ze worden ook nader ontsloten door gebruik te maken van bestaande, uitgebreide indices die samen met de resoluties deel uitmaken van het archief. Deze worden gedigitaliseerd en getransformeerd en opgenomen in een digitaal ontsluitingsapparaat waar ze kunnen worden gecombineerd met bestaande en nieuwe sets van ontsluiting, waaronder de bestaande ontsluitingen van de gedrukte uitgaven van de resoluties, de database van ambtenaren en ambtsdragers, en via named entity recognition verkregen persoons- en instellingsnamen.
Hier is de inbreng van alle disciplines nodig om (a) de indices zodanig te modelleren dat ze optimaal toegankelijk en doorzoekbaar zijn en (b) in een systeem gecombineerd kunnen worden met de externe bronnen. Om de opbouw en de evolutie van het sociale netwerk rondom de Staten-Generaal en ook de interactie tussen politiek-institutionele structuur, sociale samenstelling op verschillende niveaus en de agency van individuen op en rondom het Binnenhof te onderzoeken is voorts een tool nodig die het mogelijk maakt personen, netwerken, onderlinge relaties en de dynamiek daarin te bestuderen. De bestaande visualisatie-instrumenten kunnen nog niet adequaat omgaan met het temporele aspect. Onderzoek, IT en DDB zullen gezamenlijk te te zetten stappen in kaart moeten brengen en uitvoeren.
Trefwoorden: combinatie papieren bronnen; gedigitaliseerde data; visualisatie (tooling)

Migrant

In het migrant project staat het ontsluiten van de maatschappelijke dynamiek rond migratie aan de hand van het Nederlands-Australische voorbeeld centraal. Het aantal beschikbare relevante collecties is groot, internationaal en wijdverspreid en deels digitaal en deels analoog beschikbaar. Uitgangspunt is steeds geweest dat de centrale dataset waar mogelijk verrijkt zou worden met gebruikmaking van digitale middelen, maar dat waar nodig uitbreiding met de hand zou plaatsvinden. Door de combinatie van zoveel bronnen en datasets met zeer uiteenlopende achtergronden, is bronnenkritiek op alle sets van groot belang. De meeste van de data komen van erfgoedinstellingen en hebben een institutionele achtergrond; het perspectief van de migranten zelf ontbreekt hierin voor een belangrijk deel.

Als uitbreiding van de data van het migrant project wordt een uitbreiding met materiaal van migranten zelf georganiseerd. Dit is momenteel vrijwel uitsluitend in analoge vorm beschikbaar; digitalisering kan het best geschieden door leden uit de community zelf om organisatorische redenen en door hun kennis van de context. In het Huygens ING is onder meer daarvoor de ICO-app ontwikkeld, waarmee zij zelf documenten (vooral brieven, foto's en bescheiden) kunnen digitaliseren en ze direct van context kunnen voorzien door toevoeging van een aantal door de onderzoekers te definiëren metadata. Tevens is samenwerking met externe partners voorzien om de centrale dataset uit te breiden met een steekproef van na 1990 vertrokken Nederlanders, zodat ervaring kan worden opgedaan met onderzoek in en modelleren van digital born archief van overheid en migrant.
Onderzoekers werken samen met data- en ICT-specialisten om de metadata te modelleren en ze in het informatiesysteem te combineren met de reeds beschikbare data over migranten. Uit analytisch oogpunt kan hieruit de rol van sleutelfiguren uit de directe omgeving van de migranten worden opgemaakt in relatie tot aspecten van beleid en migrant agency.
Trefwoorden: modelleren; crowd-editing; onderzoek digital born archief

Boekbesprekingen
Boekbesprekingen van Nederlandse en vertaalde buitenlandse romans zijn op vele websites te vinden en bieden een beeld van de veranderende perceptie en receptie van literatuur. Het samenstellen van een representatief corpus van boekbesprekingen, om onderzoeksvragen hierover te kunnen adresseren, vergt het vergaren van de ruwe gegevens van veel verschillende websites, elk met eigen vormgeving en inhoudelijke structuur van de besprekingen, vaak met beperkte informatie over het besproken boek en de recensent. Om te achterhalen welk boek besproken wordt, moet metadata uit de webpagina worden gehaald en opgezocht worden in bibliografische databases. Recensenten geven beperkte informatie over zichzelf en zijn actief op verschillende online platforms, waardoor koppeling van verschillende besprekingen van dezelfde recensent beperkt mogelijk is. Onzekerheden omtrent identificatie en koppeling van entiteiten in het opgebouwde corpus moeten transparant weergegeven worden. De interactie tussen recensenten, in de vorm van commentaren bij elkaars bespreking en verwijzingen in eigen besprekingen naar die van andere moet ook gemodelleerd worden om de ontwikkeling van de discussie en invloeden van perspectieven te kunnen traceren. Omdat er continu nieuwe romans en besprekingen verschijnen, is het ook belangrijk om het corpus steeds uit te breiden. Versiebeheer van het corpus is belangrijk voor het rapporteren van analyses in publicaties, om herhaalbaarheid en controle mogelijk te maken, voor zover de ethische beperkingen van dit soort web data dat toestaan.
Onderzoekers werken samen met ICT-specialisten aan het identificeren van boekbesprekingen op het web, het extraheren van de aanwezige metadata en het koppelen van gegevens. Ook wordt er samengewerkt aan data mining experimenten voor het analyseren van meningen, sentiment en invloed in het netwerk van besprekingen.
Trefwoorden: webharvesting; versiebeheer; provenance; datamining

---