Strategische dataproductie: representativiteit van data via crowdsourcing

Nicoline van der Sijs (senior-onderzoeker en coördinator), Anna Kirstein (onderzoeksassistente), Daan Broeder (techniek)

december 2015

0. Introductie

Dit rapport is geschreven ten behoeve van de KNAW Agenda Grootschalige Onderzoeksfaciliteiten. Een belangrijke doelstelling van die agenda is dat digitale datasets van teksten in 2025 representatief zijn en niet langer biased. Dit rapport beschrijft wat er nodig is om die doelstelling te halen. Daartoe hebben we drie enquêtes afgenomen. Aan onderzoekers hebben we gevraagd welke lacunes in de data en metadata hun onderzoek belemmeren (§ 1.1 en bijlage 1). Ter vergelijking hebben we gesprekken gevoerd met de grootste digitaliseringsorganisaties over hun digitaliseringsprogramma's (§ 1.2). Om te beoordelen of dataproductie middels crowdsourcing haalbaar is, hebben we aan projectleiders van crowdsourcingprojecten gevraagd naar hun bevindingen (§ 2.1 en bijlage 2), en hebben we bij vrijwilligers geïnventariseerd wat hun ervaringen en wensen zijn (§ 2.2 en bijlage 3).

Op basis van de antwoorden hebben we vastgesteld hoe de crowd het best kan worden ingezet voor dataproductie (§ 3) en wat de eisen zijn voor een crowdsourcing-infrastructuur (§ 4, bijlage 4 en 5). We eindigen met enkele opmerkingen over IPR- en privacy-issues (§ 5). Aan het slot van iedere paragraaf formuleren we aanbevelingen voor de KNAW-toekomstagenda.

1. Lacunes in data en metadata

1.1 Enquête voor onderzoekers

Via een enquête hebben we bij onderzoekers van verschillende geesteswetenschappelijke disciplines gepeild waar de lacunes voor hun onderzoek liggen, wat hun toekomstige behoeftes en prioriteiten aan data en metadata zijn en wat ze verwachten als opbrengst van crowdsourcing. De enquête (bijlage 1) is ingevuld door 59 onderzoekers, verspreid over verschillende geesteswetenschappelijke disciplines en gespecialiseerd in verschillende periodes, van wie twee derde werkzaam is in Nederland. De meeste onderzoekers houden zich bezig met gedrukte of handgeschreven teksten, voornamelijk Nederlandstalige.

De grootste behoefte bestaat, zo bleek uit de vragen 1 t/m 13, aan diplomatische of kritische transcripties van teksten of gesproken taal; voor 44 onderzoekers is automatisch lezen (ocr) onvoldoende. Voor 16 onderzoekers, veelal gespecialiseerd in de 20ste en 21ste eeuw, zijn teksten gelezen met ocr voldoende. Bijna 60% van de onderzoekers heeft voldoende aan een vaste, kleine set metadata (auteur, titel, jaar en plaats), terwijl ruim een derde behoefte heeft aan meer metadata, van verschillende aard, zoals sociolinguïstische informatie over sprekers of schrijvers, algemene informatie over de bewaarplaats, genre, musicologische informatie.

Op de vraag welke lacunes in de data een belemmering voor het onderzoek vormen, zijn gevarieerde, soms gedetailleerde antwoorden gekomen. Vrij veel respondenten wijzen in het algemeen op grote behoefte aan een goed uitgebalanceerd (qua tekstsoorten en periodes) historisch en hedendaags tekstcorpus verrijkt met taalkundige en syntactische informatie, dat goed doorzoekbaar is ook voor taalkundige onderzoeksvragen, met betrouwbare tekstdateringen en transcripties, met

gebruiksvriendelijke interface, mogelijkheden om deelcorpora samen te stellen, gemakkelijk exporteerbaar naar een dataverwerkingsprogramma, ook exporteerbaar als platte tekst en doorzoekbaar met reguliere expressies.

Gedetailleerdere wensen van de onderzoekers betreffen de volgende datasets:

1. Middeleeuwse en vroegmoderne handschriften en incunabels

("zowel kleurenfoto als transcriptie", "middeleeuwse oorkondes uit Brabant en Vlaanderen"; "gelokaliseerde teksten"; "corpora uit de 15e en 16e eeuw ("devotionele, theologische en wetenschappelijke teksten"; "digitale edities zijn gebaseerd op onbetrouwbare 19e-eeuwse uitgaven of bevatten te veel ocr-fouten")

2. Diachroon corpus egodocumenten

(" met de nadruk op de 15e, 16e en 17e eeuw")

3. Dialectcorpora

("lacunes voor de 18e en 19e eeuw, gedifferentieerd per dialectregio of -locatie"; "moderne dialectgegevens: samenvoegen van de databases van semasiologische en onomasiologische dialectwoordenboeken; transcripties van dialectbanden van Meertens Instituut en UGent ; transcripties van de Reeks Nederlandse Dialectatlassen")

4. *Meer genres*

("er is een bias richting canon"; "almanakken, kroniekjes, goedkoop drukwerk, volksboekjes, dagboeken"; "preken"; "nonfictie")

5. *Meer literaire tijdschriften en kranten van de 20ste eeuw*

("De Stem"'; "kranten t.o.v. verzuiling"; "kranten uit de koloniën")

6. Gedrukte officiële besluiten

("resoluties Staten-Generaal, Raad van State (ancien regime); staatsbladen, staatscouranten vanaf 1813; Nederlandse Jurisprudentie")

7. Lexicale data

("neologismen"; "de resources achter de lexicografische datasets van bv. INL dringend openbaar beschikbaar worden (i.e. de databanken moeten vrij downloadbaar worden), om bestaande systemen voor woordherkenning / lemmatisering / spellingnormalisatie in alle openheid te kunnen verbeteren")

8. Archieven

("NSB-archief, Centraal Archief Bijzondere Rechtspleging")

9. Andere talen dan het Nederlands

("brieven ed. in het Engels geschreven door Nederlanders"; "Neolatijn uit de renaissance"; "talen in de Democratische Republiek Congo"; "ook Fries en Afrikaans")

10. Spraakcorpora

("met 'rijke' annotaties, zowel van het Nederlands als van andere talen"; "alle TV-programma's met ondertiteling"; "geluidsopnames met transcripties van regionale variatie"; "opnames van (NL) sprekers met een dysartrie - spraakproblemen die voortkomen uit neurologische aandoeningen of schade")

11. Sociale media en webteksten

("geschreven socialemedia- teksten in niet-Standaardnederlands"; "verrijkte Corpora from the Web")

12. Videodata

("van natuurlijke interactie van verschillende alledaagse en institutionele interacties van (near) native sprekers van het Nederlands, bijv. gezin in de keuken, rijles, afrekenen in een winkel, klasseinteractie, medische interactie etc.")

1.2 Plannen van de Koninklijke Bibliotheek en Metamorfoze

Om te achterhalen in hoeverre al in reguliere digitaliseringsplannen wordt tegemoet gekomen aan wensen van de onderzoekers, hebben we overlegd met vertegenwoordigers van de Koninklijke Bibliotheek en Metamorfoze (Steven Claeyssens, Jasper Faase, Marg van der Burgh). Hieruit bleek dat volgende aantallen boeken in samenwerking met diverse partners zijn gedigitaliseerd; deze titels zijn of komen beschikbaar in Delpher:[1]

Periode	Collectie	Aantal	Partners/project
1500-1900	Google-project KB, UBA	332.479 banden	Samenwerking Google
1781-1800	EDBO	10.000 banden	Samenwerking UBL, UBA, KB
1900-1909	BNB 2 KB, UBA, UBU	16.000 banden	Boekentraject Metamorfoze
1900-1919	BNB Bezen 2 UBL	4.537 banden	Boekentraject Metamorfoze
1910-1919	BNB 3 KB	4.065 banden	Boekentraject Metamorfoze
1920-1939	BNB 4, 5, 6 unieke titels KB	30.000 banden	Boekentraject Metamorfoze
1618-1995	kranten	8.000.000 pag.	KB
19e-20e eeuw	tijdschriften	1.500.000 pag.	KB
1937-1989	radiobulletins	1.800.000 pag.	ANP

Het is de ambitie van KB is om in 2030 de complete eigen collecties aan gedrukte boeken, tijdschriften, kranten en (middeleeuwse) handschriften gedigitaliseerd te hebben, dat wil zeggen gescand en waar mogelijk gelezen met optische tekenherkenning. Tot 2018 worden via verschillende routes nog zo'n 150.000 extra boeken gedigitaliseerd. Daarnaast wordt voor 2018 de digitalisering van krantentitels van nationaal belang afgerond. Welke titels van nationaal belang wordt afgestemd met de wetenschappelijke adviescommissie van bureau Metamorfoze. Momenteel is slechts 10% van de kranten gescand.

In samenwerking met ProQuest zijn 18.206 banden met handschriften tot 1700 uit de KB-collectie gedigitaliseerd hebben (maar deze zijn momenteel alleen via ProQuest beschikbaar).[2] Recent is KB zelf begonnen met digitalisering van 78.995 handschriften uit de collectie van na 1700, maar de prioriteit ligt bij digitalisering van boeken, kranten en tijdschriften.

Metamorfoze zal de komende jaren veel handgeschreven archieven digitaliseren, dat wil zeggen scannen, niet machinaal lezen. De grootste collectie bestaat uit 100 dozen (overeenkomend met ca. 160.000 scans) gekaapte brieven uit het National Archive in Kew; deze komen in 2017 beschikbaar, maar zonder metadata en zonder transcriptie.

De overlap tussen de wensen van onderzoekers en wat sowieso door KB gedigitaliseerd wordt, is op titelniveau op dit moment niet vast te stellen. Het is echter duidelijk dat de digitalisering van de Koninklijke Bibliotheek en andere nationale erfgoedinstellingen tot nu toe nog niet heeft geleid tot een representatieve set van gedigitaliseerde teksten, omdat de digitalisering van teksten in het Metamorfoze-programma zich vooral richt op gedrukte materialen uit de moderne periode (vooral 1840-1950), dat onderhevig is aan verval. De digitalisering van middeleeuwse manuscripten en archiefstukken is grotendeels overgelaten aan het initiatief van individuele erfgoedinstellingen. Deze materialen zijn uniek, en van groot belang voor de Nederlandse taal en geschiedenis. Een bijkomend probleem is dat de teksten zich niet alleen bevinden in Nederland en België, maar verspreid over de wereld bewaard worden.

Voorts voldoet optische tekenherkenning van oudere teksten niet aan de wensen

[1] http://www.delpher.nl/

[2] http://eeb.chadwyck.co.uk/geoLocSubscription.do

van de meeste onderzoekers; hier zal kwaliteitsverhoging moeten plaatsvinden via crowdsourcing. Het is namelijk gebleken dat crowdsourcing een effectieve methode is om slechte ocr te corrigeren en verrijken. Daarbij is KB zeer geïnteresseerd in concrete samenwerking met KNAW op dit terrein, omdat ook KB kwaliteitsverhoging nastreeft; er bestaan al samenwerkingsverbanden tussen KB enerzijds en Meertens Instituut en NIOD anderzijds op het gebied van 17de-eeuwse kranten respectievelijk verzetskranten uit de Tweede Wereldoorlog. Voor publicaties uit de 21ste eeuw geldt geen kwaliteitsprobleem want de politiek van KB is om deze zoveel mogelijk direct in digitale vorm op te nemen (hier geldt wel het probleem van IPR, zie § 5).

1.3 Aanbevelingen voor het dichten van lacunes in data en metadata

Om zinnige uitspraken te kunnen doen over lacunes in de data en om een concrete digitaliseringsagenda te kunnen opstellen, bevelen wij aan te investeren in een project dat als doel heeft een gedetailleerd overzicht te vervaardigen over de omvang van het gedrukte en handgeschreven Nederlandstalige erfgoed van 1000 tot heden, wat daarvan bewaard is gebleven, en wat daarvan is gedigitaliseerd of binnenkort gedigitaliseerd zal worden. [3] Het Nederlandstalige erfgoed bestaat uit middeleeuwse handgeschreven manuscripten en incunabelen (tot 1470), gedrukte publicaties (1470 tot heden) en archieven (veelal unieke teksten, deels handgeschreven deels getypt). Van alle drie deze teksttypen moeten inventarisaties worden gemaakt. De vraag naar de omvang van de Nederlandse productie (inclusief grijze literatuur) dient getrapt onderzocht te worden.

1. Wij bevelen aan project op te starten voor het inventariseren van het gedrukte Nederlandstalige erfgoed, door:

a) bestaande digitale catalogi van gedrukte publicaties te koppelen en te ontdubbelen : NCC, Picarta, STCN, Nederlab-auteurscatalogus, Nationale Thesaurus voor Auteursnamen (NTA, http://www.oclc.org/support/ services/ggc/nta.en.html), VIAF (Virtual International Authority File, http://viaf.org/) en CLARIN VLO;

b) deze gegevens te vergelijken met bestaande gedrukte overzichten zoals Saalmink, de Brinkman e.d., waarvan de inhoud moet worden omgezet naar structurele data;

c) deze gegevens te vergelijken met boekhoudingen van uitgeverijen/boekhandels/etc. of studies op basis van dergelijke archiefstukken.

Alleen zo kunnen we gedegen uitspraken doen over de omvang van de productie door de eeuwen heen, het aandeel dat nog in papieren vorm tot ons gekomen is, het aandeel dat veel of weinig gelezen werd en het aandeel dat (tussentijds) digitaal beschikbaar is.

Op basis van de aldus samengestelde duurzame digitale thesaurus van Nederlandse gedrukte teksten kunnen door onderzoekers gerichte digitaliseringsverzoeken worden gedaan op basis van concrete onderzoeksvragen, waarna in overleg met KB en Metamorfoze een digitaliseringsprogramma kan worden opgezet om die lacunes te dichten.

2. Wij bevelen aan project op te starten voor het inventariseren van de Nederlandse manuscripten van vóór 1501, zowel binnen als buiten Nederland, inclusief alle overgeleverde fragmenten. Daarvoor moet het bestaande overzicht van middeleeuwse handschriften (Bibliotheca Neerlandica Manuscripta) worden voltooid en bij de tijd gebracht. Hierbij dient te worden opgemerkt dat juist middeleeuwse handschriften

[3] In het beleidsplan 2010-2013 van KB wordt de boek-, krant- en tijdschriftproductie in Nederland geschat op 700 miljoen pagina's, zie https://www.kb.nl/sites/default/files/docs/beleidsplan_kb_1013.pdf, p. 6

veelal ontbreken in de bestaande gedigitaliseerde tekstcollecties, ofwel alleen zijn opgenomen in een latere, gedrukte editie.

3. Wij bevelen aan een project op te starten voor het inventariseren van de inhoud van Nederlandstalige archieven, zowel binnen als buiten Nederland, gebaseerd op bestaande overzichten van archieven (zoals: Archiefnet, Archieven.nl, Archive Grid) en egodocumenten (zoals Catalogus Epistularum Neerlandicarum; R. Dekker e.a., *Egodocumenten van Noord-Nederlanders van de zestiende tot begin negentiende eeuw. Een chronologische lijst*, Rotterdam 1993; R. Lindeman e.a., *Reisverslagen van Noord-Nederlanders van de zestiende tot begin negentiende eeuw: een chronologische lijst,* Haarlem 1994). Vooronderzoek is gedaan door DEN Enumerate; naar schatting is momenteel circa 10% van de archieven gedigitaliseerd (lees: gescand - dus niet getranscribeerd).[4] Op basis van de inventarisatie kan een stappenplan en prioritering voor toekomstige digitalisering worden opgesteld in gezamenlijk overleg tussen archieven, onderzoekers en Metamorfoze.

4. Wij bevelen aan voor de ontsluiting van audiovisuele gegevens - waaraan binnen de geënquêteerde groep slechts beperkte behoefte blijkt - samenwerking te zoeken met daarin gespecialiseerde partners zoals de groep die voor de KNAW- toekomstagenda de ADVANT (ADvanced Video Analysis Tool) uitwerkt.

5. Gezien de te verwachten enorme toename van digital born-teksten en scans bevelen wij aan in overleg met grote collectiebeheerders als KB en met archieven (NIOD, IISG, NA, Netwerk Digitaal Erfgoed Decentraal) afspraken en richtlijnen op te stellen voor formaten en minimale metadata van digital born-teksten; momenteel bestaan hiervoor geen algemene richtlijnen, waardoor het linken van digital born-teksten aan elkaar en aan gedigitaliseerde oudere tekstcollecties problematisch is en het onderzoek belemmert. Een eerste stap is gezet in de Nationale strategie digitaal erfgoed: http://www.den.nl/pagina/511/netwerk-digitaal-erfgoed/.u

6. Wij bevelen aan na te denken over vervolgprojecten van de infrastructuurprojecten Nederlab en CLARIAH om tegemoet te komen aan de breedgevoelde wens naar een uitgebalanceerd verrijkt diachroon tekstcorpus. Beide infrastructuurprojecten voorzien in generieke zoekapplicaties en generieke taggers en parsers. Binnen een vervolgproject moeten gespecialiseerde taggers en parsers worden ontwikkeld voor semi-automatische verrijking van een representatief diachroon corpus met handmatige correctie, zodat een duurzaam onderzoeksinstrumentarium beschikbaar komt, als diachrone aanvulling op onder andere het Corpus Gesproken Nederlands.

2. Ervaringen met crowdsourcing

Om erachter te komen of crowdsourcing daadwerkelijk een zinnige methode is om te komen tot dataproductie, hebben we enquêtes afgenomen bij projectleiders en vrijwilligers.

2.1 Ervaringen van projectleiders

Van de onderzoekers heeft 13% enige ervaring met crowdsourcing (bijlage 1, vraag 20 en 21). De meesten zijn daarover enthousiast: "leverde veel data en geen problemen op",

[4] Zie: http://www.den.nl/art/uploads/files/DEN/Enumerate-core-survey-NL2012-2013-def20131017-CC.pdf

maar ook: "De vrijwilligers moeten goed opgeleid en begeleid worden. Ook zijn sociale activiteiten gewenst om hen gemotiveerd te houden. Men heeft soms de neiging om daarop te besparen, wat echter contra-productief is."[5]

Tot slot hebben we 7 ervaren projectleiders, van grote en kleine projecten, gevraagd naar hun ervaringen; zie bijlage 2. De projecten behelsden voornamelijk transcriberen en metadateren.

De projectleiding/coördinatie nam per project veelal gemiddeld een dag per week in beslag, bij kleine projecten wat minder maar niet heel veel minder, omdat in die gevallen de projectleiding vaak inhoudelijk meewerkt aan het vrijwilligersproject. Er werd geen eigen technische voorziening gebouwd maar gewerkt met bestaande software/tools, wat de meeste projectleiders betreuren omdat ze daardoor tegen allerlei technische beperkingen oplopen. Er wordt gewerkt met gedetailleerde instructies. Voor de kwaliteitscontrole worden speciale vrijwilligers ingezet, soms doet de projectleider hier ook aan mee. Er waren verschillende beloningssystemen. Geen van de projecten maakt gebruik van gaming. Werving gebeurde via de media of via persoonlijke contacten.

Alle projectleiders zijn tevreden over de resultaten. De meeste projecten hebben een budget van een paar duizend euro; het grootste project had ongeveer € 50.000 aan manuren, dat niet alleen voor crowdsourcing werd toegepast maar ook voor scannen, databeheer en dergelijke. Dit grootste project werkte via VeleHanden van Picturae, en betaalde € 16.000 voor het gebruik van de website voor een periode van negen maanden. Conclusies van de projectleider van dit project zijn relevant voor toekomstplannen (meer in bijlage 2): “Het project was opgezet als experiment in het kader van een groter digitaliseringsprogramma. We wilden onderzoeken in hoeverre crowdsourcing financieel en kwalitatief een meerwaarde heeft in het digitaliseren van collectiedata. Die meerwaarde is zeker gebleken.”

Lessen die de projectleiders noemen zijn: “Denk niet al te veel na in naam van de crowd”; “vertrouw op je crowd”; “geef goede instructie en feedback”; “zorg voor persoonlijk contact”.

De belangrijkste gemeenschappelijke ervaring is dat er zich aan het begin vaak veel mensen aanmelden, waarvan er ook heel snel weer redelijk wat afvallen. Daarna blijft er een harde kern over die gestaag doorwerkt. Over het algemeen wordt ongeveer 90 % van het werk gedaan door 10% van de vrijwilligers. Toch draagt iedereen bij, en geen van de projectleiders vindt dat er een selectie van vrijwilligers dient plaats te vinden (behalve op inhoudelijke criteria: als het werk voor een vrijwilliger te hoog is gegrepen).

2.2 Ervaringen van de crowd

Door de Stichting Vrijwilligersnetwerk Nederlandse Taal en het Meertens Instituut zijn 5 vrijwilligersprojecten georganiseerd. Aan alle vrijwilligers van deze projecten, ook die zijn gestopt, is een enquête verzonden om te achterhalen wat hun ervaringen zijn. Er zijn 247 antwoorden binnengekomen in 2 weken, zie bijlage 3.

Op basis van de antwoorden kunnen we het volgende profiel van de vrijwilligers vaststellen:

- Gemiddelde leeftijd is 63,4 jaar; de jongste vrijwilliger is 23, de oudste 99.
- 51,6% is vrouw, 48,4% is man.
- 80,5% heeft een hbo- of academische opleiding genoten; de beroepen zijn uiterst

[5] Van onderzoekster Montserrat Prats Lopez, die in 2016 hoopt te promoveren op *Citizen Science: A multiple case study in the humanities* bij de VU, hebben we informatie gekregen over de ervaringen met andere crowdsourcingprojecten en relevante literatuur daarover (verwerkt in bijlage 4 en 6).

gevarieerd maar conform de hoge vooropleidingen.
- Ruim 37% van de vrijwilligers heeft aan meer dan één project meegewerkt, en 66,5% van de vrijwilligers geeft op dat ze ook ander vrijwilligerswerk doen of gedaan hebben (de aard van dat andere vrijwilligerswerk vertoont enorme variatie).
- De meeste vrijwilligers hebben over het project gehoord via de media of nieuwsbrieven.
- Redenen om mee te doen waren in aflopende volgorde: de wetenschap helpen; nieuwe kennis en ervaring opdoen of belangstelling voor het onderwerp. Sociale contacten, carrière en waardering spelen nauwelijks een rol.
- 50,8% werkt nog mee (sommigen sinds 2007), 38,7% is definitief gestopt, de overigen hebben veelal een tijdelijke pauze ingelast; de redenen om te stoppen waren voornamelijk tijdsgebrek, persoonlijke omstandigheden, gebrek aan persoonlijk contact, minder interessant, project klaar en technische problemen.
- De beoordeling over de begeleiding, instructies, projectleiding, contacten met andere vrijwilligers en waardering zijn voor het merendeel goed of voldoende. Daarbij valt op dat er vrij vaak wordt geantwoord dat men geen behoefte heeft aan contact met andere vrijwilligers.
- Contacten met de projectleiding en andere vrijwilligers worden onderhouden via e-mail, Yahoo-group of Forum. Ruim 68% heeft rechtstreeks contacten per e-mail met de projectleiding (van wie sommigen expliciet melden dat zij daar de voorkeur aan geven) en geen contacten met andere vrijwilligers, 17% maakt gebruik van een Forum en bijna 7% is lid van een Yahoo-group.
- De meeste vrijwilligers vinden de taken niet te moeilijk en geven de voorkeur aan pittige opdrachten omdat makkelijke opdrachten te saai zijn.
- 56% van de vrijwilligers ziet niets in het toevoegen van spelelementen (gaming) bij crowdsourcing, bijna 30 % heeft geen mening, en slechts 6,5% vindt dit een goed idee. Uit de antwoorden blijkt dat de meeste vrijwilligers óf geen óf een vrij stereotiep-negatieve voorstelling hebben van spelelementen ('kinderachtig', 'onvolwassen gedoe', 'niet meer serieus te nemen'). De meeste vrijwilligers willen absoluut geen competitief element in het werk. Beloningen in de vorm van bijeenkomsten of cadeautjes worden wel gewaardeerd.

De algehele indruk is dat de vrijwilligers serieuze, hoogopgeleide en oudere personen zijn die vrijwilligerswerk doen om zich maatschappelijk nuttig te maken of als zinnige tijdsbesteding. Ze hebben weinig behoefte aan sociale contacten met andere vrijwilligers, maar wel aan rechtstreeks contact met de projectleiding.

2.3 Aanbevelingen naar aanleiding van opgedane ervaringen

6. Wij bevelen op basis van de opgedane ervaringen aan om strategische dataproductie via crowdsourcing te laten verlopen en daarvoor een professionele organisatie op te zetten, een KNAW crowdsourcingplatform.

7. Wij bevelen aan een campagne te starten om de wetenschappelijke werkzaamheden van verschillende geesteswetenschappelijke en levenswetenschappelijke KNAW-instituten bekend te maken en het publiek te vragen mee te werken als vrijwilliger of als informant. Hiervoor moet een vragenlijst worden opgesteld die aansluit op de CLARIAH Online Survey tool die momenteel in ontwikkeling is. Op basis van de ingevulde vragenlijsten wordt een vrijwilligersregister samengesteld. Uit het register kunnen vrijwilligers met geschikte profielen worden gelicht voor specifieke projecten of taken, maar daarnaast moeten de vrijwilligers de vrijheid behouden zelf keuzes te maken, omdat zij dat waarderen, zo bleek uit de enquêtes.

3. De inzet van de crowd

3.1 Centraal platform

De behoefte van onderzoekers aan nieuwe data of aan verbetering van bestaande data blijkt groot, wat automatisch betekent dat er veel vrijwilligers nodig zijn, en diverse specifieke vrijwilligersprojecten. Uit de ervaring van andere crowdsourcingprojecten (zie bijlage 4) blijkt dat de oprichting van een centraal platform voor crowdsourcing het beste werkt om massa te mobiliseren. Op dit platform kunnen veel verschillende projecten worden getoond, wat een aanzuigende werking op vrijwilligers heeft. De KNAW kan zich daarbij profileren ten opzichte van andere crowdsourcingplatforms (zoals VeleHanden) door zich te concentreren op data voor wetenschappelijk onderzoek, door hoge kwaliteit en flexibiliteit (volgens een geënquêteerde projectleider problematisch bij VeleHanden). De oprichting van een platform vergt een eenmalige investering, maar levert voordelen door schaalvergroting.

3.2. Taken van de crowd

Uit de wensen van de onderzoekers naar data en metadata kunnen we de taken van de crowd globaal in twee typen verdelen:

- ontsluiten van bestaande teksten, beelden, gesproken taal door middel van transcriptie of metadata (§ 1, 1 t/m 9);
- verzamelen van nieuwe, experimentele data of toevoegen van annotaties (§ 1, 10 t/m 12).

Deze taken vergen verschillende crowds. Voor het ontsluiten van materiaal is een crowd nodig waarover veel informatie bekend is (zoals opleiding en ervaring), die langere tijd meewerkt aan het project, en die, in ieder geval voor een deel, specialistische kennis heeft. Voor deze groep, met wie vaak een lange relatie wordt opgebouwd, is een vrijwilligersadministratie nodig, met inlogprocedure en langdurige opslag van gebruikersinformatie. Het verzamelen van nieuwe, experimentele data vereist vaak inzet van een grote diverse groep mensen van wie weinig informatie beschikbaar hoeft te zijn.

De meeste geënquêteerde onderzoekers willen de crowd met name inzetten omdat de technische mogelijkheden van automatische tekst- en spraakherkenning momenteel onvoldoende zijn. Sommigen zeggen dat hun complexe en specifieke data (bijv. regionaalgekleurd taalgebruik) niet geschikt zijn voor niet-menselijke ontsluiting. Als belangrijkste concrete taken van de crowd (zie bijlage 1, vragen 14 t/m 18) zien onderzoekers het transcriberen van teksten, fouten in de ocr corrigeren, en metadata of annotaties laten toevoegen aan teksten of bestaande metadatasets laten uitbreiden. Voor taalkundige verrijking heeft men met name behoefte aan lemmatisering, Parts-Of-Speech-tagging, syntactische parsing en Named Entity Recognition.

32 onderzoekers denken dat het nuttig zou zijn studenten of jonge onderzoekers voor crowdsourcing in te zetten. Dit lijkt een verantwoordelijkheid van de onderzoekers zelf, die hiervoor een beloningssysteem kunnen afspreken met hun studenten. 28 onderzoekers verwachten geen specifieke achtergrondkennis bij de crowd, wat de werving vereenvoudigt. Daarentegen hebben 23 onderzoekers vrijwilligers met gespecialiseerde kennis nodig, van (Middeleeuwse) letterkunde, ouder Nederlands, paleografie, taalkunde of Latijn.

Gezien de grote hoeveelheid data die onderzoekers ontsloten willen zien, zijn wij van oordeel dat de meest efficiënte werkwijze bestaat uit een combinatie van technische oplossingen met crowdsourcing. Een veelbelovende ontwikkeling, die meer mogelijkheden biedt dan GIWIS (Groningen Intelligent Writer Identification System) is

het lopende project Transkribus,[6] waarbinnen automatische handschriftherkenning, gebaseerd op een lexicon, wordt gecombineerd met handmatige correctie door vrijwilligers.

3.3 Gaming

Uit de reacties van de vrijwilligers bleek dat de meerderheid niets ziet in het toevoegen van spelelementen of gaming. Waarschijnlijk is hier sprake van onbekendheid met het fenomeen en met bias, doordat de geënquêteerde vrijwilligers zich bezighouden met transcriberen. Dat werk vinden velen op zich al een spannend spel, en hieraan hoeft geen gaming te worden toegevoegd. Daarentegen blijkt gaming wel erg geschikt voor het verrijken of annoteren van teksten en het verzamelen van experimentele data, zo tonen in Groningen ontwikkelde games voor Engelstalige teksten aan;[7] annotaties kunnen ook worden verzameld via het opensource CrowdTruth Framework.[8]

3.4 Aanbevelingen voor de inzet van de crowd

9. Wij bevelen aan crowdsourcing te combineren met technische oplossingen; in concreto adviseren wij een tool als Transkribus voor semi-automatische handschriftherkenning verder te ontwikkelen voor Nederlandstalige handgeschreven en gedrukte (oude, jonge, in Gotisch schrift gedrukte) teksten, en hiervoor Nederlandse lexica (van INL of automatisch gegenereerde) en applicaties als TICCL en PICCL in te zetten.

10. Wij bevelen aan gaming-software te ontwikkelen als hulpmiddel voor het verrijken van Nederlandstalige teksten en het verzamelen van experimentele data.

4. Benodigde crowdsourcing-infrastructuur

4.1 Organisatie

Het onderhouden van een crowdsourcingplatform vergt structurele tijdinzet. Uit de enquêtes van projectleiders en vrijwilligers komt naar voren dat communicatie met de vrijwilligers een belangrijk aspect vormt van de motivatie en waardering: zonder persoonlijk contact valt een groot aantal vrijwilligers vrij snel af. Voor grootschalige crowdsourcing is in onze ogen na de initiële investering structureel aan personeelskosten nodig: 0,2 algemene projectleider/senior onderzoeker, 0,2 technicus voor onderhoud van de software en oplossen van problemen, en 0,6 vrijwilligers-manager.

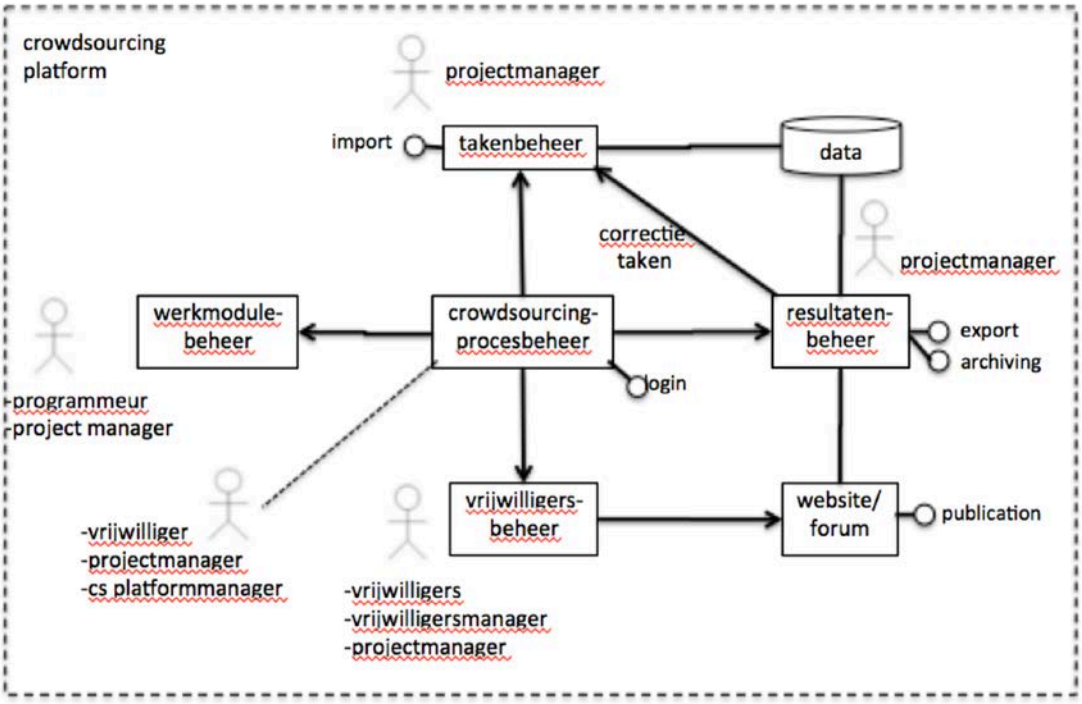
4.2. Technische infrastructuur

Voor crowdsourcing is een infrastructuur nodig die veel verschillende taken kan regelen, met een modulaire en schaalbare opzet; voor een gedetailleerdere uitwerking zie bijlage 5. Aangezien het op dit moment onbekend is welke middelen kunnen worden aangewend om een infrastructuur te realiseren noch wanneer zo'n taak van start gaat, schetsen we de infrastructuur alleen in algemene termen. De infrastructuur bestaat uit een webapplicatie/website, omdat die technisch het best kan worden onderhouden.

Essentieel voor succesvolle crowdsourcing is een laagdrempelige technische voorziening, een tool die de taken van de crowd in behapbare porties aanbiedt en zoveel

[6] https://transkribus.eu/Transkribus/
[7] http://wordrobe.housing.rug.nl/Wordrobe/public/HomePage.aspx
[8] http://crowdtruth.org/

mogelijk autonoom werkt. De tool ondersteunt het werk van de crowd en dat van de organisatie en geeft voortdurend feedback over de voortgang van de werkzaamheden.



De crowdsourcing-infrastructuur bevat vijf onderdelen, verantwoordelijk voor:
a. *vrijwilligersbeheer*, voor de regeling van de vrijwilligersadministratie: aanmelden, inloggen, editen, uitschrijven, rollenbeheer, communicatie (Forum, e-mail, nieuwsbrief, Q&A, FAQ, chatbox);
b. *crowdsourcingprocesbeheer*, voor het maken van rapportages over vrijwilligers, taken en resultaten ten behoeve van de projectleider en het toewijzen van taken aan gebruikers;
c. *takenbeheer*, voor de administratie van taken: importeren, specificeren, editen, toebedelen, verwijderen van taken;
d. *werkmodulebeheer*, voor het inpluggen van specifieke modules, bijvoorbeeld een module voor semi-automatische correctie van ocr;
e. *resultatenbeheer*, voor het beheren van de geproduceerde resultaten.

De crowdsourcing-infrastructuur moet het mogelijk maken een speciale groep daarvoor geëquipeerde vrijwilligers te laten optreden als corrector. Uit de uitgevoerde enquêtes blijkt dat zowel projectleiders als vrijwilligers er behoefte aan hebben dat ingevoerde correcties voor anderen zichtbaar zijn, zoals in een Wiki: op die manier leren de vrijwilligers van gemaakte fouten en is het proces voor iedereen inzichtelijk.

4.3 Aanbevelingen voor de crowdsourcing-infrastructuur

11. Wij bevelen aan structureel voor crowdsourcing een budget uit te trekken van 0,2 algemene projectleiding, 0,4 techniek en 0,6 vrijwilligersmanager.

12. Wij bevelen aan zowel voor de crowdsourcing-infrastructuur als voor de werkmodules en het platform eerst bestaande applicaties te inventariseren en te testen voordat er iets nieuws wordt gebouwd. Daarvoor moet eerst een gedetailleerd model met gewenste functies en functionaliteiten worden vastgesteld.

5. IPR- en Privacy-issues

Voor de toekomstagenda zijn goede afspraken rond IPR en privacy essentieel: zonder dat is onderzoek niet reproduceerbaar en controleerbaar. 40% van de onderzoekers rapporteert problemen te hebben met IPR en privacy. Indirect ervaren nog veel meer onderzoekers hiermee problemen, want het gebruik van grote collecties als van Koninklijke Bibliotheek en universiteitsbibliotheken, en van grote infrastructuur-projecten is aan allerlei IPR-beperkingen onderworpen. KB zet zich in om afspraken over IPR te maken met auteursrechtelijke organisaties, en komt daarin ook geleidelijk aan steeds verder. KNAW zou hierin samen met KB en andere nationale instellingen kunnen optrekken, want het is niet verstandig aparte afspraken te maken.

Voor privacy-gevoelig materiaal zijn we gehouden aan de Wet Bescherming Persoonsgegevens. Die wet geldt ook voor de samenstelling van een centraal vrijwilligersregister: aan de vrijwilligers moet toestemming worden gevraagd hun persoonsgegevens te gebruiken en bekend te maken op bijvoorbeeld een forum. Vrijwilligers moeten voor aanvang van werkzaamheden informatie krijgen over hun rechten en plichten, en hen moet worden gevraagd het auteursrecht op transcripties en dergelijke over te dragen of vooraf toestemming te verstrekken voor het gebruik van de transcripties, bijvoorbeeld: “Iedereen heeft en houdt de vrije beschikking over het zelf ingetikte gedeelte. Medewerking houdt in dat het resultaat van het werk om niet en belangeloos beschikbaar wordt gesteld voor wetenschappelijk onderzoek, bij voorkeur op een algemeen toegankelijke website. De namen van de vrijwilligers/medewerkers worden daarbij expliciet vermeld, behalve wanneer een vrijwilliger dat niet wenst. ”

5.1 Aanbevelingen rond IPR en privacy

13. Wij bevelen de KNAW aan om, eventueel in gezamenlijkheid met andere betrokken instellingen, te komen tot specifieke richtlijnen voor IPR en privacy met betrekking tot crowdsourcingprojecten, en met collectiebeheerders gemeenschappelijk op te trekken voor zaken rond IPR op data.