Bijlage 1:

Enquête crowdsourcing KNAW voor onderzoekers en collectiebeheerders

Algemene informatie:

59 reacties in 4 weken

Geachte collega's,

Binnen de geesteswetenschappen van de KNAW ontwikkelt een consortium momenteel plannen om een onderzoeksagenda voor de toekomst (2025) te formuleren, getiteld DESIDERIA (Dutch Extensible and Searchable Infrastructure for Digital Explorative Reading and Information Analysis). Die agenda is gericht op onderzoek aan de hand van grote tekstcorpora en sluit aan bij huidige (inter)nationale projecten als CLARIN en CLARIAH. De agenda bouwt voort op bestaande projecten door op drie cruciale punten naar vooruitgang in reeds aanwezige tekstuele digitale infrastructuren te streven. Het accent ligt hierbij op infrastructuren rond Nederlandstalig materiaal, materiaal dat zich voor een belangrijk deel in Nederland bevindt, en taalbeschrijvingen die door Nederlandse onderzoekers zijn gemaakt.
Een van de drie actiepunten betreft het onderzoeken van welke lacunes er zijn in bestaande datasets die bias veroorzaken met betrekking tot de mogelijke representativiteit van de genereerde data waardoor vernieuwend onderzoek kan worden belemmerd, en hoe die lacunes met behulp van crowdsourcing kunnen worden opgelost. Het consortium heeft ons verzocht een rapport te schrijven met een draaiboek voor het crowdsourcen van wetenschappelijke data. Hiervoor doen we een beroep op u.
Wij verzoeken u zo veel mogelijk vragen van de onderstaande vragenlijst te beantwoorden. Als u verder nog suggesties heeft of reeds ervaring met eigen crowdsourcingprojecten heeft opgedaan, zijn wij hier zeer benieuwd naar. Voor vragen en opmerkingen kunt u terecht bij Anna Kirstein (anna.kirstein@meertens.knaw.nl) of Nicoline van der Sijs (post@nicolinevdsijs.nl).
Alvast hartelijk dank voor uw medewerking.

Met vriendelijke groet,

Nicoline van der Sijs (coördinatie)
Anna Kirstein (onderzoeksassistentie)
Daan Broeder (techniek)

Autobiografische informatie

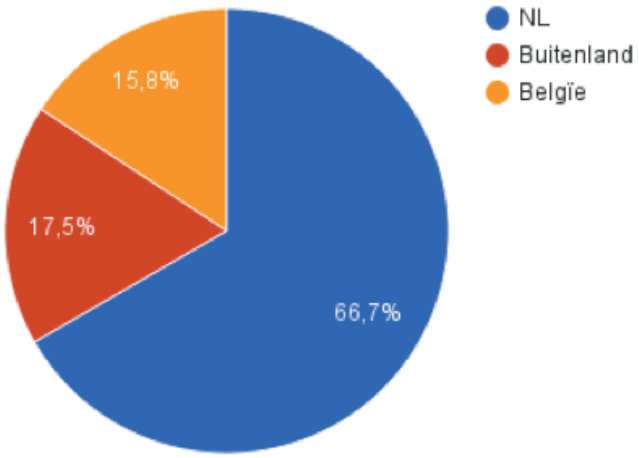
Voor de verdere verwerking van de enquête hebben wij enige informatie van de invullers nodig. Deze informatie zal alleen gebruikt worden als onderlegger voor het rapport dat ten behoeve van de KNAW wordt opgesteld. In dit rapport zullen alle gegevens worden geanonimiseerd.

1. Achternaam en initialen: Er zijn 59 reacties binnengekomen.

2. E-mailadres

3. Aan welke instelling(en) bent u verbonden?

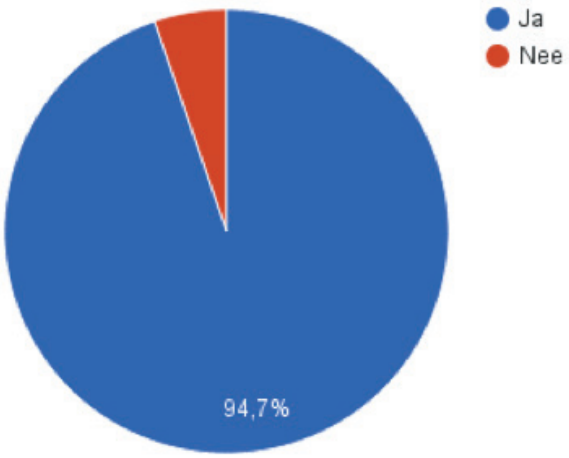
Instelling uit



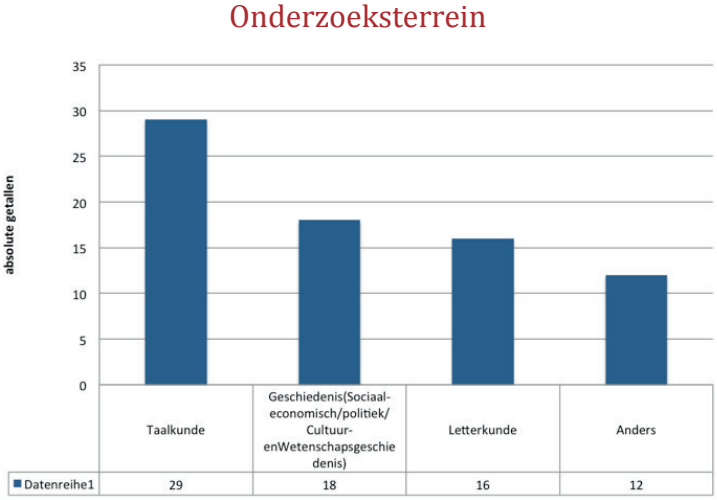
4. Welke functie bekleedt u?

5. Wilt u op de hoogte gehouden worden van het onderzoek?

Wilt u op de hoogte gehouden worden van het onderzoek

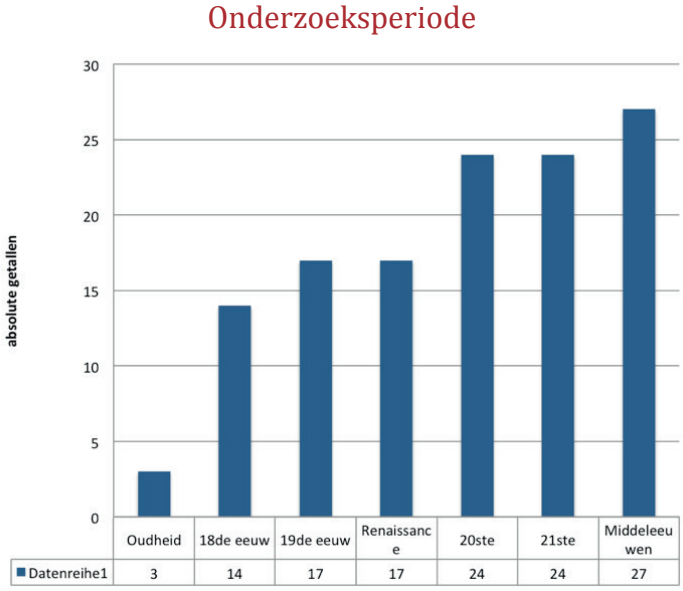


6. Op welk terrein verricht u onderzoek? (Meerdere antwoorden mogelijk, staafdiagram geeft absolute aantallen)



a. Taalkunde: 29
b. Geschiedenis: 18
c. Letterkunde: 16
d. Anders (rechtsgeschiedenis, spraakpathologie, taal- en spraaktechnologie, cultuurwetenschap, boekwetenschap, antropologie, codicologie/paleografie, e-humanities, klassieke talen, kunstgeschiedenis): 12

7. Uit welke periode? (meerdere antwoorden mogelijk, staafdiagram geeft absolute aantallen)



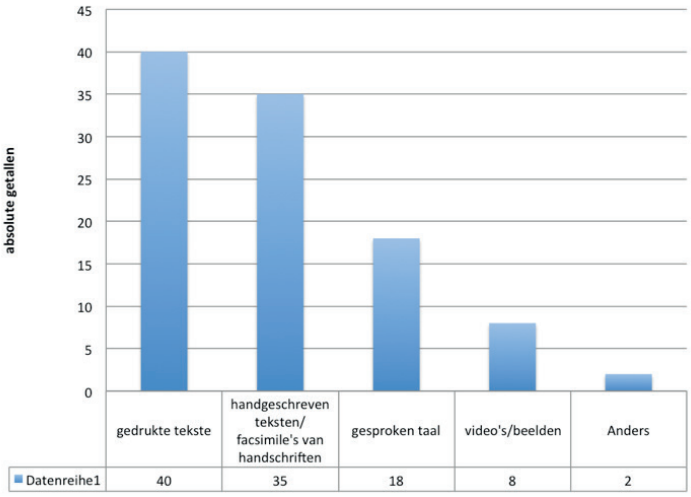
a. Middeleeuwen: 27
b. 20ste eeuw: 24
c. 21ste eeuw: 24
d. Renaissance: 17
e. 19e eeuw: 17
f. 18de eeuw: 14
g. Oudheid: 3

8. Welke lacunes in de data vormen een belemmering voor uw onderzoek? [open vraag, geen grafiek]

Vragen m.b.t. de gewenste onderzoeksdata

9. Welk formaat hebben de teksten/data die belangrijk zijn voor uw onderzoek? (meerdere antwoorden mogelijk, staafdiagram geeft absolute aantallen)

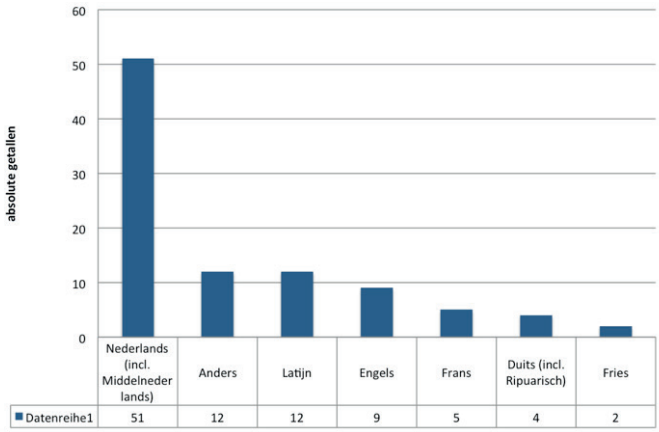
Lacunes formaat



a. gedrukte teksten: 40
b. handgeschreven teksten/ facsimile's van handschriften: 35
c. gesproken taal: 18
d. video's/beelden: 8
e. Anders: 2

10. In welke taal/talen is/zijn de bronnen geschreven? (meerdere antwoorden mogelijk, staafdiagram geeft absolute aantallen)

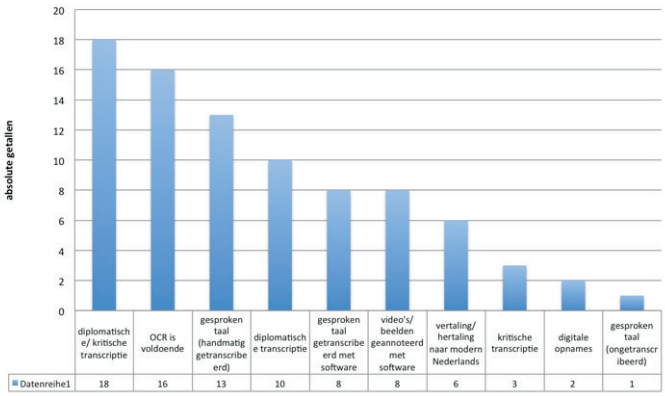
Talen



- Nederlands: 51
- Latijn: 12
- Engels: 9
- Frans: 5
- Duits: 4
- Fries: 2
- Anders (Afrikaanse, Aziatische, Latijns-Amerikaanse, moderne westerse, West-Europese talen, Hanzegebiedtalen, West-Germaans): 12

11. Hoe zouden de data eruit moeten zien? (meerdere antwoorden mogelijk, staafdiagram geeft absolute aantallen)

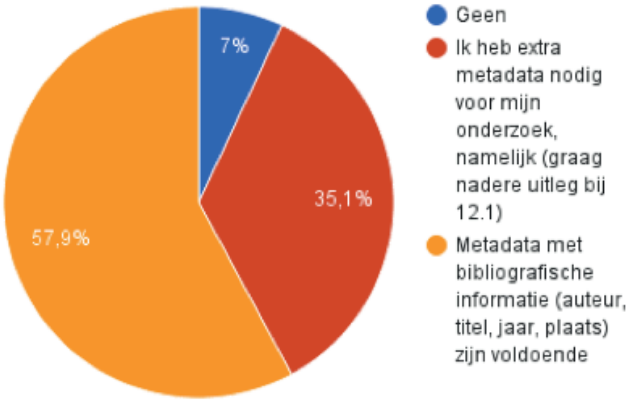
Gewenste data



- diplomatische of kritische transcriptie zijn allebei geschikt: 18
- gelezen met optische tekenherkenning is voldoende: 16
- gesproken taal handmatig getranscribeerd: 13
- diplomatische transcriptie: 10
- gesproken taal getranscribeerd met PRAAT of een andere software: 8
- video‘s/ beelden geannoteerd met ELAN of een andere software: 8
- vertaling/ hertaling naar modern Nederlands: 6
- kritische transcriptie: 3
- Anders: digitale opnames (2), gesproken taal ongetranscribeerd (1)

12. Welke metadata heeft u nodig bij uw data?

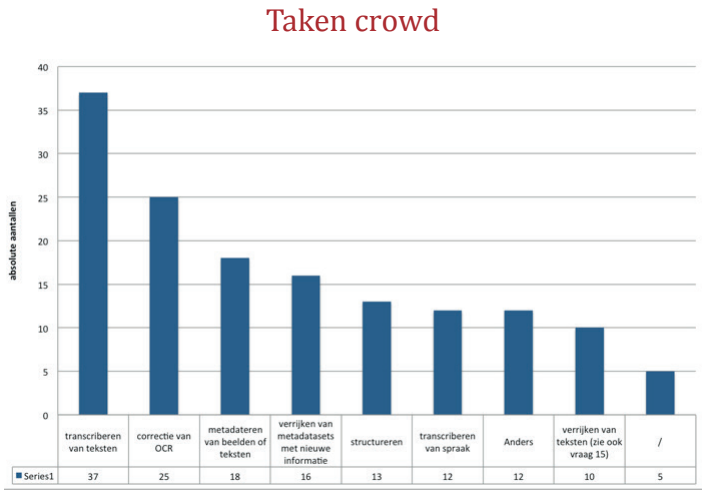
Gewenste metadata



1. Metadata met bibliografische informatie (auteur, titel, jaar, plaats) is voldoende: 57,9%
2. Extra metadata nodig: 35,1% (sociolinguïstische informatie over sprekers/ schrijvers, algemene informatie over bewaarplaats, tekstsoort, musicologische informatie, etc.)
3. Geen metadata nodig: 7%

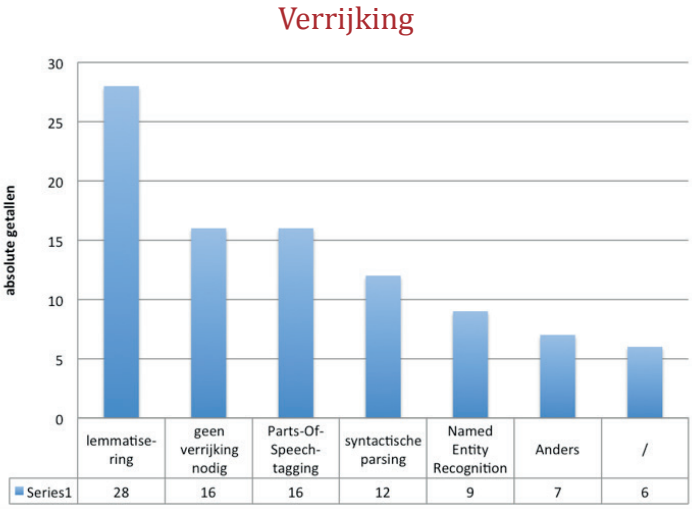
13. Zijn er lacunes in de metadata die een belemmering vormen voor uw onderzoek? Welke? [Open vraag, geen grafiek]

14. Voor welke doeleinden zou u de ‚crowd' willen inzetten m.b.t. uw data? (meerdere antwoorden mogelijk, staafdiagram geeft absolute aantallen)



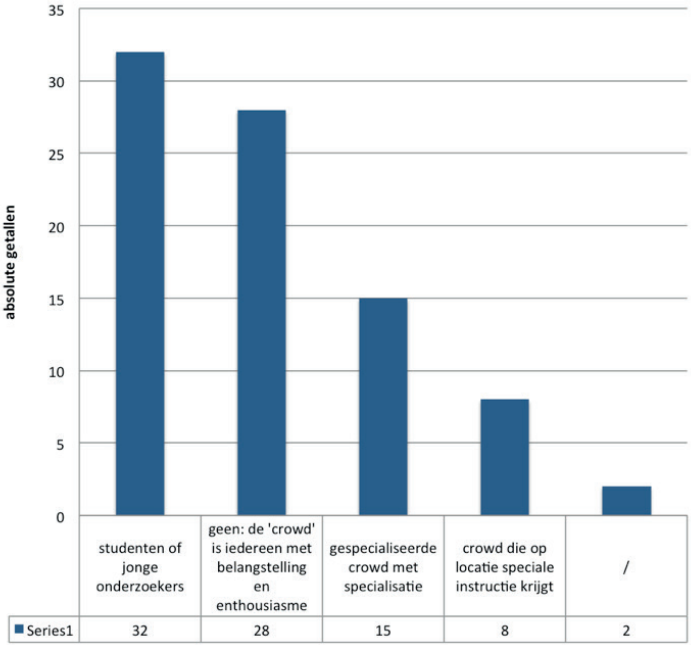
- transcriberen van teksten: 37
- corrigeren van fouten in de optische tekenherkenning: 25
- metadateren van beelden of teksten: 18
- verrijken van metadatasets met nieuwe informatie: 16
- ongestructureerde data (teksten) omzetten naar gestructureerde (database) (bijvoorbeeld gegevens uit handboeken en encyclopedieën omzetten naar database formaat, of sociaal-economische data interpreteren in database...): 13
- transcriberen van spraak: 12
- verrijken van teksten (zie ook vraag 15): 10
- Anders (digitaliseren, correctie automatische spraakherkenning, verstaanbaarheid, oordelen over (uit)spraak, mate van correctheid, projectmanagement): 12
- *Geen mening*: 5

15. Hoe zouden uw data verrijkt moeten worden? (meerdere antwoorden mogelijk, staafdiagram geeft absolute aantallen)



- lemmatisering: 28
- Parts-Of-Speech-tagging: 16
- Verrijking is voor mijn onderzoek niet nodig: 16
- syntactische parsing: 12
- Named Entity Recognition: 9
- specifieke verrijking (transcriptie met woordgrenzen, mate van correctheid, oordelen over (uit)spraak, semantic tagging, thematische parsing): 7
- *Geen mening*: 6

16. Wat verwacht u aan achtergrondkennis van de ‚crowd' die uw (meta)data kan ontsluiten? (meerdere antwoorden mogelijk, staafdiagram geeft absolute aantallen)



- studenten of jonge onderzoekers: 32
- Geen: de ‘crowd’ is iedereen met belangstelling en enthousiasme voor mijn onderwerp: 28
- gespecialiseerde crowd met specialisatie: 15 (nadere uitleg bij volgende vraag)
- crowd die op locatie speciale instructie krijgt: 8
- *Geen mening*: 2

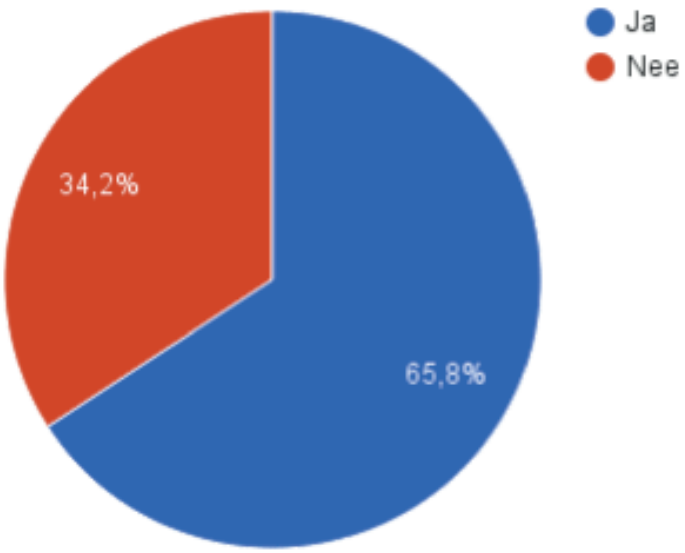
16.1 Specialisatie van de ‚crowd‘ (optioneel, open vraag)

Reacties van 19 onderzoekers, antwoorden ingedeeld in categorieën

- historische kennis van (Middeleeuwse) taal- en letterkunde en het (Middel) Nederlands
- specifieke kennis van een taal/dialect
- fonetische of taalkundige kennis
- kennis van paleografie/codicologie
- Latijn

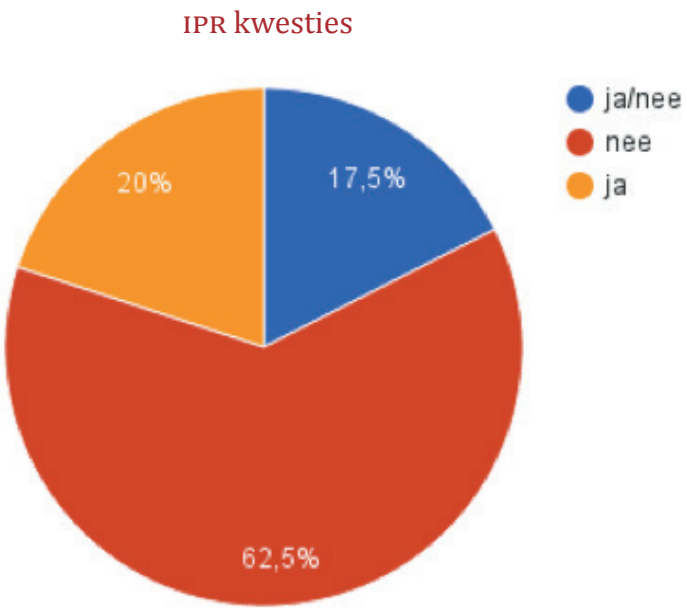
17. Zou u ook overwegen om betaalde krachten te gebruiken voor het ontsluiten van uw data? Wat zijn uw overwegingen dit niet/wel te doen?

Totaal inzet betaalde krachten



18. Moeten uw wensen door de crowd worden uitgevoerd of kan een verbeterde techniek aan uw wensen tegemoet komen? [Open vraag, geen grafiek]

19. Ondervindt u bij uw onderzoek problemen m.b.t. de copyright die op teksten berust? Wat zijn die problemen?



nee: 62,5% (niet van toepassing omdat de teksten oud)
ja: 20%
geen probleem met IPR, wel met de privacy-wetgeving: 17,5%

20. Heeft u voor eerder onderzoek reeds gebruik gemaakt van crowdsourcing?

    13%: ja
    Welke problemen heeft u daarbij ondervonden? En welke voordelen heeft u ervan gehad? [open vraag, geen grafiek]

21. Kent u goede voorbeelden van crowdsourcing waar we ons voordeel mee kunnen doen? [open vraag, geen grafiek]

22. Verdere opmerkingen/ suggesties/ ideeën: [open vraag, geen grafiek]