# Het Woogle Woo-dossiercorpus

Maarten Marx, Maik Larooij, Guido Enthoven & Jaap Kamps[1]

**De Woo is nu anderhalf jaar van kracht en leidt tot steeds meer actieve publicatie van documenten door de bestuursorganen die onder de Woo vallen. Met Woogle worden die documenten herbruikbaar. Documenten van een breed palet aan bestuursorganen worden verzameld en voorzien van uitgebreide gestandaardiseerde metadata.**

Woogle[2] is een platform ontwikkeld aan de Universiteit van Amsterdam dat beoogt op één plek alle documenten die zijn vrijgegeven onder de Wet Open Overheid (Woo) vanuit alle bestuursorganen die onder de Woo vallen samen te brengen. Woogle is hierin vergelijkbaar met het *informatieregister*,[3] uit een eerdere versie van de Woo, wat later afgezwakt werd tot het Platform Open Overheidsinformatie (PLOOI), dat na het BIT-advies van eind 2022[4] nog verder is afgezwakt tot de Woo-index. De term 'informatieregister' lijkt tijdens de behandeling van de Woo besmet geraakt. Dit kwam doordat het idee nooit is uitgewerkt en een centrale en directe toegang tot alle documenten van de hele overheid als onhaalbaar en zeer kostbaar is ingeschat.[5]

De Woo is nu anderhalf jaar van kracht en moet leiden tot steeds meer actieve publicatie van documenten door de bestuursorganen die onder de Woo vallen.

## Woogle en herbruikbaarheid

Een belangrijk doel achter Woogle is om die documenten *herbruikbaar* te maken, vooral voor de wetenschap, en in het bijzonder voor grootschalig computerondersteund diachronisch comparatief onderzoek. Het verzamelen, preprocessen, normaliseren en harmoniseren van data neemt vaak tot wel 80% van de onderzoekstijd in beslag en vergt, zeker bij data op grote schaal, technische vaardigheden die juristen en sociale wetenschappers meestal niet in huis hebben. Binnen Woogle is dit allemaal al gedaan: documenten van een breed pallet aan bestuursorganen zijn op allerlei plaatsen verzameld, ze krijgen uitgebreide gestandaardiseerde metadata, alle tekst wordt machineleesbaar gemaakt en alles is beschikbaar in een open machine lees- en bewerkbaar formaat, via DANS Easy.[6] Hier staat een bevroren 'dump' uit mei 2023, en hyperlinks naar dumps die elke nacht bijgewerkt worden. De hyperlinks naar de originele PDF-versies van de vrijgegeven documenten staan in de metadata. Eigenlijk zijn die PDF's alleen nodig voor kwaliteitscontrole of handmatige analyse. Soms zijn de stukken niet meer beschikbaar omdat ze zijn weggehaald of verplaatst. In dat geval zijn de originele PDF's via de Woogle site beschikbaar.

## Woo-dossiers in Woogle

Een afgehandeld Woo-verzoek bestaat uit het verzoek, het besluit, en mits aan het verzoek voldaan is, een inventarislijst van de relevante stukken en natuurlijk alle vrijgegeven stukken. We noemen dit een Woo-dossier. Begin oktober 2023 bevatte Woogle 9181 Woo-dossiers. Daarin zitten respectievelijk 7955, 2387, 1497 en 36.073 losse besluiten, verzoeken, inventarislijsten en bestanden met vrijgegeven stukken. Het is zeer lastig iets te zeggen over het aantal vrijgegeven documenten omdat die in de regel niet als losse documenten maar aan elkaar geplakt in één (vaak enorm groot) PDF-bestand zitten. In totaal gaat het om meer dan 1,2 miljoen bladzijden in 48.000 bestanden. De door Woogle via optische karakterherkenning (OCR) uit de PDF's verkregen teksten bestaan uit 339 miljoen woorden uit een vocabulaire van 5,9 miljoen unieke woor-

**De enorme verbeterslag gemaakt door de extra OCR blijkt als we dit vergelijken met wat er in de oorspronkelijke bestanden zat**

© Shutterstock

den, waarvan er 3,5 miljoen slechts éénmaal voorkomen. De enorme verbeterslag gemaakt door de extra OCR blijkt als we dit vergelijken met wat er in de oorspronkelijke bestanden zat: 28 miljoen woorden uit een vocabulaire van 6,3 miljoen woorden, met meer dan 4 miljoen hapaxen. Met de extra OCR van goede kwaliteit hebben we meer dan tien keer zoveel data, en de kwaliteit is beter want er zijn minder OCR-verhaspelingen. De collectie komt echt op gang na 2018, met de volgende aantallen dossiers per jaar: 2023: 1629; 2022: 2348; 2021: 2057; 2020: 1848; 2019: 840; 2018: 191. De dossiers komen uit dertien gemeentes, tien provincies, twaalf ministeries en een klein aantal ZBO's. De dossiers zijn vaak slecht terug te vinden via Google en ook via het eigen Woo-platform van de overheid[7] door de gebrekkige kwaliteit van de documenten en de indexerings- en OCR-strategie van Google. Omdat Woogle elk document opnieuw verwerkt met de krachtige Tesseract OCR, is de (terug)vindbaarheid daar een stuk beter.

## Mogelijke onderzoeksvragen

We geven wat voorbeelden van onderzoeksvragen over de besluiten na Woo-verzoeken die met het Woogle-corpus te beantwoorden zijn.

- Hoe vaak worden de verschillende weigeringsgronden echt toegepast? We kunnen dit tellen per dossier, per vrijgegeven document, en zelfs per bladzijde.
- Wat staat er in de inventarislijsten en wat is de kwaliteit daarvan? Hoe uniform zijn die lijsten binnen een bestuursorgaan en over organisaties?[8]
- Hoe, en hoe uitgebreid wordt een besluit gemotiveerd?
- Voldoen de openbaar gemaakte Woo-dossiers aan de Wet Hergebruik Overheidsinformatie[9] en aan artikel 2.4 lid 3 van de Woo?[10]

Door de uniforme opzet van het Woogle-corpus zijn comparatieve en/of diachrone versies van deze vragen ook makkelijk te beantwoorden. •

**Auteurs**

1. Dr. M. Marx, M. Larooij BSc, Faculteit der Natuurwetenschappen, Wiskunde en Informatica, Universiteit van Amsterdam; dr. G. Enthoven, Instituut voor Maatschappelijke Innovatie; dr. J. Kamps, Faculteit der Geesteswetenschappen, Universiteit van Amsterdam.

**Noten**

2. woogle.wooverheid.nl.
3. *Kamerstukken II* 2015/16, 33328, nr. 32.
4. adviescollegeicttoetsing.nl/documenten/publicaties/2022/11/28/bit-advies-plooi.
5. Advies ACOI Juli 2023; acoi.nl/publicaties/publicatie/brief-aan-de-minister-van-bzk.
6. doi.org/10.17026/dans-zau-e3rk.
7. open.overheid.nl.
8. M. Larooij et al., *ESB* 2023, 108(4817), p. 36-39, esb.nu/036-039_larooijkamp-2.
9. wetten.overheid.nl/BWBR0036795/2021-07-01.-
10. wetten.overheid.nl/jci1.3:c:BWBR0045754&hoofdstuk=2&artikel=2.4&z=2023-04-01&g=2023-04-01.