ESB draft Fair Wob dossiers

- marx
- 2022-04-28
- `pandoc --filter pandoc-citeproc --bibliography="ESB draft Fair Wob dossiers.bib" -s "ESB draft Fair Wob dossiers.md" -t latex -o "ESB draft Fair Wob dossiers.tex"`
- ESB eisen aan artikelen

Titel: WOB dossiers voldoen niet aan FAIR data principes

Maik Larooij, Maarten de Rijke, Jaap Kamps, Maarten Marx

Recent onderzoek van Open State [@openstate2022] keek naar de *tijd* die nodig is om een Wob (Wet Openbaarheid Bestuur, sinds 1 Mei Wet Open Overheid (WOO)) besluit te nemen, en maakte een schatting van de kosten per pagina. Gemiddeld kost het 161 dagen om aan een verzoek te voldoen, en komt het in 80% later dan de maximale termijn. Open State schatte de kosten op 150 Euro per pagina. In het onlangs verschenen jaarlijkse rapport van *Reporters sans frontières (RSF)* is Nederland gezakt van de 6e naar de 28e plaats op het gebied van persvrijheid. Een van de genoemde redenen is dat door journalisten via de Wob opgevraagde (documenten te laat, niet correct of onvolledig loskomen (Voetnoot naar https://rsf.org/en/country/netherlands).

Wij kijken hier naar de *kwaliteit* van de gepubliceerde Wob besluiten, inclusief de vrijgegeven documenten. We kijken bewust niet naar de inhoud van de stukken en dus niet naar de vraag of de informatiebehoefte gesteld in het Wob verzoek wel adequaat is vervuld. De kwaliteit wordt beoordeeld op basis van de 4 FAIR principes van open data en voor overheidsinformatie in het bijzonder [@wilkinson2016fair]. Deze zeggen dat data vindbaar, archiveerbaar, uitwisselbaar en herbruikbaar moet zijn.

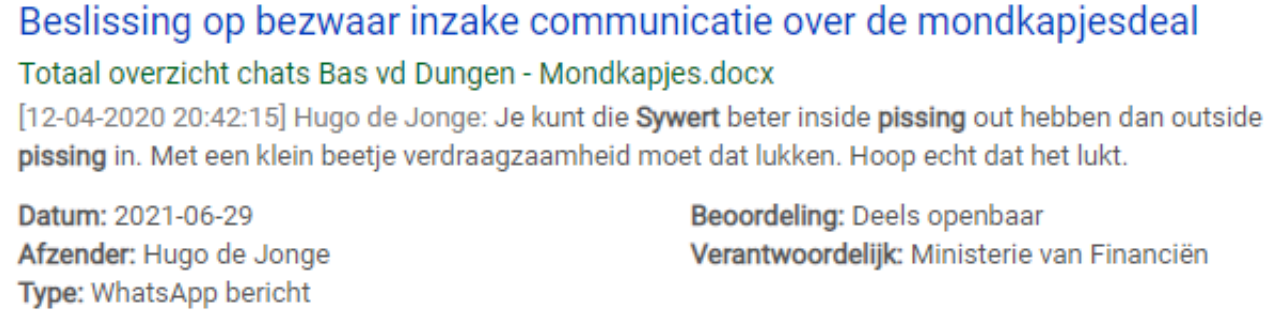
We willen dus de volgende onderzoeksvraag beantwoorden:

> *Hoe staat het met de FAIRness van de door de Nederlandse overheid gepubliceerde Wob dossiers?*

We zullen zien dat dat niet best gesteld is, maar we laten ook zien dat deze situatie heel eenvoudig, tegen vrijwel nihile kosten, en met een enorm potentieel aan opbrengsten omgedraaid kan worden.

De 4 principes zijn nogal abstract en niet eenvoudig te operationaliseren. Daarom beantwoorden we de vraag via een gedachtenexperiment. We stellen ons voor dat we een zoekmachine, een gespecialiseerde Google, voor Wob dossiers gaan maken. Vanzelfsprekend spelen de 4 principes dan een grote rol. Om de werking concreet te maken geven we een voorbeeld.

Stel een onderzoeker wil de exacte zin hebben die Minister de Jonge schreef over Sywert van Lienden en plassen in de tent. Na een zoekvraag "sywert pissing" hoopt zij dan een antwoord te krijgen op de volgende Google-achtige manier:



Een zoekmachine voor Wob dossiers

De vrijgegeven Wob dossiers worden steeds beter toegankelijk gemaakt, zeker door de centrale overheid, dus we kunnen aannemen dat we die allemaal beschikbaar hebben. Om een zoekmachine te maken moeten die dossiers wel aan de volgende 3 basisvoorwaarden voldoen:

1. *De logische informatieeenheid komt overeen met de technische bestandseenheid.* Google leidt je naar een Wikipedia pagina, niet naar de hele Wikipedia. Een verwijzing naar een citaat uit de Bijbel is altijd zo precies mogelijk. Het technische formaat van Wikipedia, elk lemma is een aparte pagina op het web, en van de Bijbel, elke zin heeft een unieke code vergelijkbaar met een URL, maakt dit mogelijk.
2. *De woorden in de documenten zijn als woorden leesbaar door een computer.* Een situatie waarin dit niet het geval is is wanneer men met Control F zoekt naar een woord in een PDF file en niks vindt terwijl dat woord toch duidelijk op het scherm staat. Het tekstbestand is dan feitelijk opgeslagen als een foto.
3. *Per informatiedrager is een zekere minimale hoeveelheid metadata aanwezig.* Documenten vindbaar op Google hebben een titel, meestal een datum, een adres (de URL) en nog veel meer metadata die de zoekmachine gebruikt om de resultaten op relevantie te ordenen en te presenteren.

We gaan deze voorwaarden nu stuk voor stuk langs.

1. *Vrijgegeven Wob documenten voldoen in overgrote meerderheid niet aan de eerste voorwaarde.* De essentie van een Wob verzoek is dat het vraagt om de *documenten* over een bepaald onderwerp. Het Wob dossier bestaat in de regel uit 3 *PDF files*, het besluit, de inventarislijst, en een PDF met daarin alle vrijgegeven documenten achter elkaar geplakt, zonder voor de computer leesbare grenzen. Alle door Open State bekeken Wob dossiers hebben deze vorm, en vrijwel ook alle dossiers vindbaar op het web die gepubliceerd zijn door lagere overheden. De gemeente Amsterdam en de provincie Gelderland vormen twee uitzonderingen. Zij plaatsen de verzameling documenten op een heel logische wijze in een zip bestand.
   Wij hebben de best beschikbare AI technieken op basis van *machine learning* met neurale netwerken toegepast op het probleem van het automatisch weer opdelen in de oorspronkelijke documenten. Dit is eigenlijk niet foutloos te doen. Onze best behaalde score had een pakkans van slechts 50% waarbij dan nog 1 op de 5 voorspelde documenten niet correct waren.
2. Veel van de Wob documenten zijn *scans* van een print. De meeste scanners staan standaard zo ingesteld dat ze *optische karakter herkenning* toepassen en niet alleen een foto maken maar de tekst ook voor een computer leesbaar maken. Bij de Wob documenten gaat dit heel vaak niet goed.
   De vrijgegeven wob dossiers op `wobcovid19.rijksoverheid.nl` geven een goede indruk van dit probleem. Toen wij ze daar ophaalden bevatten die 28.331 paginas. Bijna een kwart daarvan bevatte geen enkel door de computer leesbaar woord. Op 98% van die paginas stonden wel woorden. De OCR leverde meer dan een miljoen extra herkende woorden op. Op 77% van de paginas vond de OCR extra informatie die daarvoor niet computer leesbaar was (en dus niet met Control F gevonden had kunnen worden).

3. De inventarislijst die bij bijna elke Wob dossier zit is een tabel met op elke rij een document, en per kolom specifieke metadata voor elk document zoals de titel, het soort document (mail, whatsapp bericht, kamerstuk, etc), hoe het is vrijgegeven, de eventuele weigeringsgrond, etc. Dit klinkt ideaal, en dat zou het ook zijn als 1) die inventarislijsten als een Excel bestand openbaar gemaakt werden en niet als een uitgeprinte en weer ingescande (en vaak onleesbare) tabel, en 2) waarbij elke Wob producent consequent dezelfde namen voor de kolommen zou gebruiken, en ook consequent is in het benoemen van de waardes in de cellen, en 3) als alle Wob producenten dat op dezelfde manier zouden doen. Jammer genoeg is dat niet het geval. We hebben 2703 Wob dossiers opgehaald op `open.overheid.nl`. Bij slechts 436 konden we daarbij op basis van de bestandsnaam een inventarislijst vinden, allemaal als tabel in een PDF. Die konden 346 keer min of meer foutloos automatisch omgezet worden naar een spreadsheet. Tabel 1 toont hoe vaak we hierin basale metagegevens konden terugvinden, en op hoeveel verschillende manieren dezelfde informatie wordt weergegeven in de inventarislijsten.

Attribuut	**% aanwezig**	**Gebruikte varianten**
Doc identifier	75%	nr; nummer; volgnummer; docnr; documentnr; id
naam document	71%	document; documentnaam; titeldocument; titeldoc; onderwerp; naamdocument; titel; naam
beoordeling	67%	beoordeling; beroordeling; oordeel; beoordelingwob
weigeringsgrond	68%	weigeringsgrond; artikelwob; wob; beslissingconform; wobgrond; uitzonderingsgrond; artikel; wobartikel; weigeringsgrondwob; weigeringsgronden; lakgrond; relevantewobgronden; grond
datum	51%	datum; datumdocument
soort	21%	soort; soortdocument; type; categorie; documenttype; typedocument; soortstuk
afzender	53%	afzender; afzenders; van
ontvanger	52%	ontvanger; ontvangers; naar; aan

Tabel 1. Aanwezigheid van standaard attributen van documenten in Wob dossiers, en de gebruikte varianten om ze weer te geven (N=346).

Dit gedachtenexperiment laat zien dat het opzeten van zo'n zoekmachine, waarmee Wob dossiers dus op een uniforme wijze *archiveerbaar* en *vindbaar* en daardoor *herbruikbaar* en via hun metadata *uitwisselbaar* worden gemaakt, een enorm lastig karwei is, puur omdat de data op zo'n onhandige en niet uniforme manier wordt aangeleverd. Gelukkig kan het ook anders.

Hoe Wob dossiers FAIR te maken?

Het lijkt erop dat Wob ambtenaren hun Wob dossiers prachtig FAIR op hun eigen schijf hebben staan, maar dat er in de laatste publicatiestap iets misgaat. Want wat is er nou eigenlijk nodig om die dossiers FAIR te publiceren? De documenten *digitaal* (en dus machine leesbaar) in een (zip) mapje, en de metadata op een uniforme wijze in een spreadsheet, via unieke codes (liefst zogenaamde *permalinks* vergelijkbaar met een DOI) gekoppeld aan de losse documenten.

Het lastigste hier is dat er heel veel Wob ambtenaren zijn met ieder hun eigen werkwijze. Dus die uniforme metadata zijn een coordinatieprobleem. Wij hebben dat simpel opgelost door gratis software beschikbaar te stellen waarmee men heel handig een Wob dossier opbouwt en automatisch uniform en FAIR publiceert (Software is gratis beschikbaar op Github en kan ook direct online gebruikt worden op https://woopublish.herokuapp.com.). De toegekende metadata zijn een uitbreiding van het voorstel door Open State, de provincie Noord Holland en de VNG [@handrijking2021].
Wob ambtenaren gebruiken al speciale software voor hun Wob dossiers, vooral om semi-automatisch persoonsgegevens te

herkennen en die zwart te lakken. Het is niet nodig om hiervoor documenten te printen en in te scannen.

Samenvattend kunnen Wob dossiers doorzoekbaar en vindbaar gemaakt worden door ze machine leesbaar en met uniforme metadata in een open formaat op `open.overheid.nl` te publiceren. Uitwisselbaarheid en herbruikbaarheid kunnen eenvoudig gegarandeerd worden door iedere Wob ambtenaar dezelfde wob publicatie software te laten gebruiken.

Wat zijn de opbrengsten?

De extra kosten om de dossiers FAIR te publiceren zijn verwaarloosbaar, zeker in verhouding tot de kosten nodig voor het vinden en anonimiseren van de stukken. De opbrengsten kunnen erg hoog zijn. Berenschot adviseert om Wob documenten netjes op te slaan en voor de Wob ambtenaar vindbaar te maken zodat ze *herbruikbaar* zijn voor een nieuw Wob verzoek (en dus niet weer opnieuw geanonimiseerd hoeven te worden) [@berenschot2021]. Prachtig natuurlijk, maar de winst lijkt in het niet te vallen bij de winst die te behalen is als *burgers eenvoudig en goed in gepubliceerde Wob dossiers kunnen zoeken* en hun vraag beantwoord zien zodat ze die dus niet als Wob verzoek hoeven te stellen.

Daarnaast kunnen onderzoekscollectieven als *Follow The Money* een hoop geld en vooral frustratie *besparen*. Zij zijn een groot deel van hun tijd kwijt met het *reverse-engineeren* van de Wob dossiers, waarbij precies de 3 bovengenoemde voorwaarden het struikelblok vormen voor het opbouwen van een bruikbaar (FAIR) digitaal dossier door zo'n journalist of collectief.

Tenslotte zijn de opbrengsten van open data en standaarden in potentie immens, zoals ook aangestipt in de recente EU Data Act voorgesteld door de Europese Commissie. De Commissie verwacht dat de herbruikbaarheid van data 280 miljard Euro aan extra BNP zal opleveren in 2028.

Conclusie

De op 1 Mei 2022 ingegane opvolger van de Wob, de Wet Open Overheid (WOO), moet, als ze haar naam eer wil aandoen, een omslag maken in de manier waarop de opgevraagde stukken openbaar worden gemaakt. We hebben hier 3 knelpunten aangegeven, met meteen daarbij een simpele en vrijwel kosteloze manier om die op te lossen. De potentiele opbrengsten zijn groot, zowel economisch als maatschappelijk.

Referenties

Enthoven, Guido, Serv Wiemers, Steef den Uijl, Arjan Nouwen, Emily Kuilman, Raoul Jorissen, and Tim Vos-Goedhart. 2022. “Ondraaglijk Traag. Analyse Afhandeling Wob-Verzoeken.” https://openstate.eu/wp-content/uploads/site s/14/2022/01/Ondraaglijk-traag-280122-def.pdf.

Enthoven, G., H. Spanninga, C. Pino, and A. Spruit. 2021. “Verbeterpunten in de Informatiehuishouding Voor Een Tijdige En Kwalitatief Goede Afhandeling van Wob-Verzoeken.” https://www.informatiehuishouding.nl/binaries/info rmatiehuishouding/documenten/rapporten/2021/04/07/verbeterpunten- informatiehuishouding-wob_rddi/Rapport+Berenschot+Verbeterpunten+ IHH+ivm+goede+afhandeling+WOB-verzoeken.pdf.

Openstate, VNG, Provincie Noord Holland. 2021. “Handreiking Open Wob. Wob-Informatie Publiceren Volgens de Wobstandaard Zodat Iedereen Er Kennis van Kan Nemen.” https://kia.pleio.nl/file/download/cd428b67-4f0b-4523-8b6a-a191e41cca80/20210420-handreiking-wob_concept_0.pdf.

Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The Fair Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3 (1): 1–9.