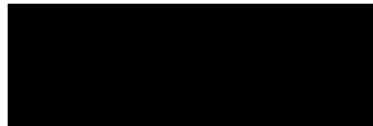


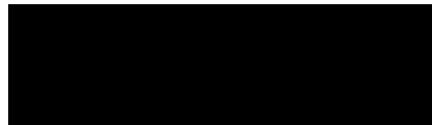
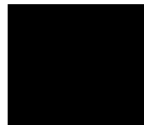
De



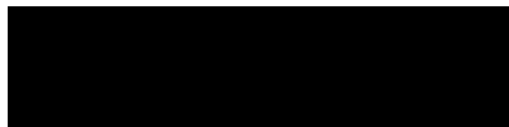
FAIRificatie



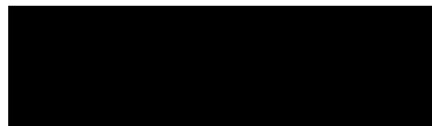
van



Woo-



dossiers



Maik Larooij

Layout: typeset by the author using L^AT_EX.
Cover illustration: Maik Larooij

De FAIRificatie van Woo-dossiers

Woo-dossiers op basis van de FAIR Data Principles voor een betere
informatiehuishouding

Maik Larooij
12761915

Bachelorscriptie
Credits: 18 EC

Bachelor *Informatiekunde*



Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam
Science Park 904
1098 XH Amsterdam

Begeleider
Dr. M.J. Marx
2e examiner
Dr. M.J. Marx

17 juni 2022

Samenvatting

TODO

Inhoudsopgave

1	Inleiding	3
2	Achtergrond	6
2.1	Wob en Woo	6
2.2	Woo-dossiers	6
2.3	FAIR Data Principles	7
2.4	Het proces van FAIRificatie	7
3	Methodologie	9
4	Pre-FAIRificatie	11
4.1	Stap 1: Analyseren van de data	11
4.2	Stap 2: Opstellen van een FAIRificatie doel	15
5	FAIRificatie	17
5.1	Stap 3: Definiëren van een semantisch datamodel	17
5.2	Stap 4: Definiëren van relevante metadata	18
5.3	Stap 5: Opstellen van beperkingen en regels	21
6	Post-FAIRificatie	24
6.1	Stap 6: Evalueren van de FAIRificatie	24
6.2	Stap 7: Het automatiseren van FAIR produceren	25
6.3	Stap 8: Identificeren van een publicatiestrategie	26
7	Conclusie en discussie	29
	Literatuur	31

1 Inleiding

De Wet openbaarheid van bestuur (Wob) is met pensioen gegaan. Op 1 mei 2022 is namelijk haar vervanger in werking getreden: de Wet open overheid (Woo). Zoals de naam ook suggereert is de wet bedoeld om de overheid transparanter te maken. De Woo regelt het recht op informatie over alles wat de overheid doet. Beide wetten stellen dat een burger een verzoek mag doen om informatie over ‘bestuurlijke aangelegenheden’. Nieuw in de Woo is de verplichting voor overheidsorganisaties om niet alleen op verzoek, maar ook zelf gefaseerd en actief informatie openbaar te maken. De Woo is hiermee een stap in de richting van een open overheid, een verschuiving die al langere tijd gaande is.

Toch lijken er ook problemen te spelen met betrekking tot het verstrekken van overheidsinformatie. Een goede en tijdige afhandeling van de verzoeken blijkt niet zo eenvoudig te zijn: het rapport en tevens oordeel ‘Ondraaglijk traag’ concludeerde dat bij ruim 80% van de (toen nog genaamde) Wob-verzoeken de wettelijke reactietermijn werd overschreden met een geschatte 150 euro aan kosten per openbaar gemaakte pagina (Enthoven et al., 2022). Eerder onderzoek aan de Technische Universiteit Delft wees in 2014 al uit dat de inhoudelijke reacties op de verzoeken zo erg van elkaar verschilden dat het zeer veel moeite kostte om een lijn in de informatie te krijgen (van Oostveen & van Loenen, 2014). Elk bestuursorgaan heeft een eigen wijze van registreren van en reageren op de informatieverzoeken. Zuiderwijk en Janssen (2014) concludeerden in een onderzoek naar het open data beleid van bestuursorganen dat er een gebrek aan samenwerking tussen de bestuursorganen lijkt te zijn, wat de informatie-uitwisseling en herbruikbaarheid van informatie belemmert. Mede door de slechte toegankelijkheid van de overheidsinformatie voor het publiek, ondanks de solide wetten, zakte Nederland het afgelopen jaar 22 plekken op het gebied van persvrijheid, van plaats 6 naar plaats 28 (Reporters Without Borders, 2022).

Verschillende initiatieven vanuit de overheid erkennen de problemen. Het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties stelde eerder al het actieplan ‘Open Overheid 2020-2022’ op, waarin wordt aangestipt dat de gepubliceerde informatie in het kader van de Wob vaak niet toegankelijk en open is (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2020). Als reactie op het rapport ‘Ongekend Onrecht’ over de Kinderopvangtoeslagenaffaire die speelde van 2004 tot 2019 stelt het kabinet dat de achterstand in de informatiehuishouding sneller moet worden ingehaald. Informatie moet volgens het kabinet, met het oog op de Archiefwet, ‘duurzaam toegankelijk, vindbaar, juist, volledig en betrouwbaar’ worden bewaard (Rutte, 2021). Berenschot stelde verbeterpunten op in de informatiehuishouding voor een tijdige en kwalitatief goede afhandeling van informatieverzoeken (Enthoven, Spanninga, Pino & Spruit, 2021). Uit dit rapport komt vooral naar voren dat de dossiers op een vindbare manier moeten worden bewaard zodat ze herbruikbaar zijn voor volgende verzoeken of openbaarmakingen.

De informatiehuishouding is niet op orde en informatie dient duidelijker, vindbaarder en beter te herbruiken te zijn. In de recente EU Data Act geeft de Europese Commissie aan te verwachten dat de herbruikbaarheid van data 280 miljard Euro aan extra BNP zal opleveren in 2028 (Europese Commissie, 2022). De betere herbruikbaarheid zou een op-

lossing moeten bieden om de verzoeken sneller te beantwoorden. Duidelijke, vindbare en uniforme data zouden de communicatie tussen de overheid en haar burgers moeten verbeteren. Kortom, een betere en toegankelijker informatiehuishouding zou een oplossing zijn voor de eerder gevonden problemen met betrekking tot de horizontale communicatie binnen de overheid en de verticale communicatie tussen de overheid en haar burgers.

Eerdere onderzoeken en rapporten stippen de noodzaak aan van een betere informatiehuishouding, maar missen echter concrete uitwerkingen van oplossingen. Enkel een conceptversie van een handreiking van de Open State Foundation in samenwerking met de Provincie Noord-Holland en de Vereniging van Nederlandse Gemeenten (2021) doet concrete aanbevelingen over metadata die opgeslagen dienen te worden bij de dossiers. De aanbeveling kan echter nog flink worden aangevuld met meer domein-relevante metadata en mist tevens nog wetenschappelijke onderbouwing. Het Forum Standaardisatie, een adviescommissie binnen het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, adviseert om aandacht te vestigen op de FAIR Principles om bij te dragen aan de ontwikkeling van concrete, herbruikbare uitwerkingen passend in het overheidsdomein (Roelfsema & de Jong, 2020). De in de wetenschap welbekende FAIR Data Principles stellen dat gegevens vindbaar, toegankelijk, uitwisselbaar en herbruikbaar openbaar dienen te worden gemaakt, precies in lijn met de wensen die in de eerder besproken actieplannen terug komen (Wilkinson et al., 2016). Het proces van het toepassen van de FAIR principes wordt in de literatuur ook wel de ‘FAIRification’ van gegevens genoemd (GO FAIR initiative, 2022; Schultes et al., 2019; Jacobsen et al., 2020). Dit onderzoek heeft als doel om op zoek te gaan naar een technische, concrete uitwerking om het bewaren en publiceren van Woo-dossiers te verbeteren aan de hand van de FAIR Data Principles. Tevens dient het als eerste verkenning van de toepassing van de FAIR Data Principles op overheidsgegevens.

Het onderzoek volgt de stappen van het proces van FAIRification. Dit beantwoordt de volgende drie onderzoeksvragen:

- 1: Hoe staat het met de *FAIRness* van de door de Nederlandse overheid gepubliceerde dossiers?
- 2: Hoe zien FAIR gepubliceerde Woo-dossiers eruit?
- 3: Hoe kan software ondersteuning bieden bij het automatisch FAIR produceren en publiceren van Woo-dossiers?

Leeswijzer

Hoofdstuk 2 verheldert eerst een aantal belangrijke concepten die van belang zijn bij dit onderzoek. Dit gaat om de Wob, de Woo, de Woo-dossiers, de FAIR Data Principles en het FAIRification proces. Hoofdstuk 3 beschrijft de methode voor het verzamelen en analyseren van data over de dossiers en gaat in op *good practices* bij het ontwerpen van software. Hoofdstuk 4 behandelt de ‘pre-FAIRification’, ofwel de noodzakelijke stappen uit het FAIRification proces voordat de daadwerkelijke FAIRification kan plaatsvinden. Hoofdstuk 5 gaat vervolgens in op de daadwerkelijke FAIRification. Hoofdstuk 6 bevat een

evaluatie van de FAIRificatie, samen met aanbevelingen om het werken met de dossiers volgens de FAIR Data Principes te bevorderen. Samen vormen deze stappen de 'post-FAIRificatie'.

2 Achtergrond

2.1 Wob en Woo

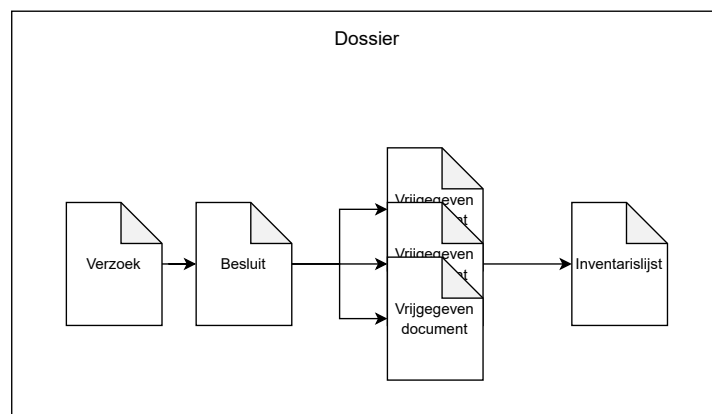
Op 1 mei 2022 is de Wet open overheid (Woo) in werking getreden. Deze wet regelt het recht van burgers op informatie van de overheid. De Woo volgt de buiten werking getreden Wet openbaarheid van bestuur (Wob) op. Artikel 3:1 van de oude Wob stelde dat “een ieder een verzoek om informatie over een bestuurlijke aangelegenheid kan richten tot een bestuursorgaan, zonder dat de verzoeker daarbij belang hoeft te hebben”. Dit recht op ‘informatie op verzoek’ is behouden in de nieuwe Woo in artikel 4:1. Een verschil in de wetten is dat de Woo ‘openbaarmaking uit eigen beweging’ verplicht. Dit betekent dat bestuursorganen worden geacht actief overheidsinformatie te publiceren zonder dat daar een verzoek voor nodig is. Naast passieve informatieverstrekking op verzoek hebben bestuursorganen er dus een extra taak bij gekregen met de ingang van de nieuwe wet: actieve informatieverstrekking uit eigen beweging.

2.2 Woo-dossiers

Documenten worden meestal openbaar gemaakt in dossiers. In dit onderzoek is een dossier gedefinieerd als ‘een collectie van openbaar gemaakte documenten naar aanleiding van een besluit op basis van de Woo’. Met de ingang van de Woo bestaan er twee soorten dossiers: dossiers op basis van een verzoek en dossiers op basis van actieve openbaarmaking. Een dossier kan bestaan uit meerdere verschillende soorten documenten. Figuur 1.2.1 laat een voorbeelddossier zien. Dit hypothetische dossier bestaat uit een verzoekdocument met daarin het oorspronkelijk verzoek van de verzoeker. Het besluitdocument bevat de reactie en beoordeling van het bestuursorgaan en geeft aan of er aan het verzoek om informatie wordt voldaan. Op basis van deze beoordeling worden er vrijgegeven documenten toegevoegd aan het dossier die de informatiebehoefte van de verzoeker moeten vervullen. Deze documenten worden vaak in tabelvorm verstrekt in een inventarislijst.

Figuur 2.2.1

Samenstelling van een dossier met in totaal zes documenten.



2.3 FAIR Data Principles

Als reactie op de noodzaak om wetenschappelijke data te kunnen hergebruiken zijn in 2016 de FAIR Data Principles geformuleerd (Wilkinson et al., 2016). De principes dienen als richtlijn bij het publiceren van data. Hoewel de principes oorspronkelijk zijn opgesteld voor wetenschappelijke data, zijn ze ook toepasbaar op andere domeinen van de samenleving waar eisen worden gesteld aan openheid en herbruikbaarheid. De verschillende principes worden hieronder op basis van de definitie van het GO FAIR initiative (2022) uitgelegd.

Findable (vindbaar)

(Meta)data moeten gemakkelijk te vinden zijn voor zowel mensen als computers. Concreet betekent dit de unieke identificatie van verschillende objecten en het toekennen van relevante metadata aan de objecten.

Accessible (toegankelijk)

De (meta)data moeten toegankelijk zijn voor zowel mensen als computers. Dit betekent dat ze op mens-, maar ook computer-leesbare en gestandaardiseerde manier moeten zijn opgeslagen en verstrekt.

Interoperable (uitwisselbaar)

De (meta)data moeten uitwisselbaar zijn. Zowel van machine op machine, van mens op mens en van machine op mens en vice versa. Dit houdt in dat de data een formele en gedeelde wijze van kennisrepresentatie nodig hebben.

Reusable (herbruikbaar)

Het belangrijkste doel van de FAIR principes is dat de data herbruikbaar zijn. Hiervoor moeten de data en relevante metadata in detail beschreven zijn en moeten ze voldoen aan domein-relevante eisen.

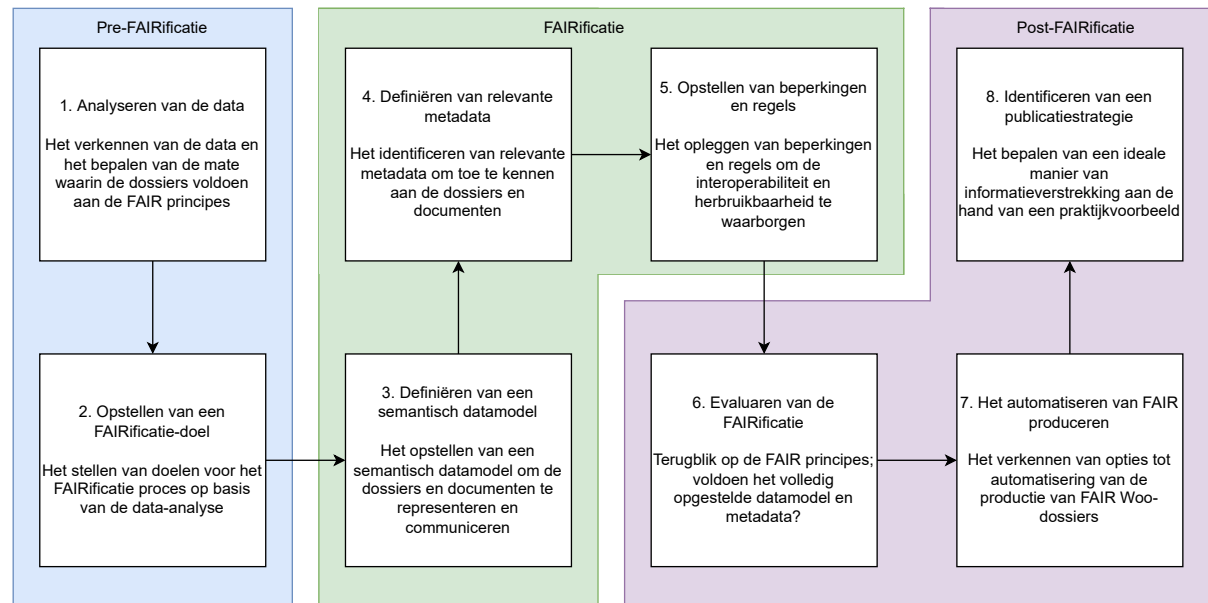
2.4 Het proces van FAIRificatie

De FAIR Data Principles beschrijven doelbewust geen technische implementatie van FAIR data. Verschillende onderzoeken zijn echter gedaan naar het proces van 'FAIRificatie', ofwel het proces van implementeren van de FAIR principes (GO FAIR initiative, 2022; Jacobsen et al., 2020; Schultes et al., 2019). Op basis van deze onderzoeken is er een eigen proces van FAIRificatie voor de Woo-dossiers samengesteld die te zien is in Figuur 2.4.1. Jacobsen et al. (2020) onderscheiden drie fases: de pre-FAIRificatie, de daadwerkelijke FAIRificatie en de post-FAIRificatie. Deze fases zijn ook terug te vinden in het opgestelde proces. In de pre-FAIRificatie worden de data geanalyseerd en worden er doelen opgesteld om te bereiken met de FAIRificatie. In de FAIRificatie fase wordt er een semantisch datamodel gedefinieerd waarin relevante metadata kunnen worden opgeslagen, samen met de in dit onderzoek toegevoegde stap van het opleggen van beperkingen en regels om de kwaliteit van de data te waarborgen. Een fase waar alle literatuur het

over eens is is de beoordelingsfase, waarin wordt geëvalueerd in hoeverre het voorstel na FAIRificatie voldoet aan de FAIR principes. Na het geëvalueerde FAIRificatie proces kan er worden gezocht naar automatische productie en uniforme publicatie van de Woo-dossiers. .

Figuur 2.4.1

Het proces van FAIRificatie voor Woo-dossiers



3 Methodologie

Om tot een antwoord te komen op de vragen hoe het staat met de FAIRness van gepubliceerde dossiers, hoe FAIR data eruit zien en hoe software ondersteuning biedt bij het FAIR produceren en publiceren van dossiers is het onderzoek opgedeeld in een aantal FAIRificatie-stappen (zie sectie 2.4). Deze sectie behandelt de dataverzameling en creatie van software.

Dataverzameling

In stap 1 heeft er een data analyse plaatsgevonden om de dossiers te toetsen op FAIRness. Hiervoor zijn twee verschillende sets aan bestanden gebruikt, te zien in Tabel 3.1.1.

Om de leesbaarheid van de documenten te toetsen zijn alle documenten uit 119 Wob-dossiers afkomstig van wobcovid19.rijksoverheid.nl gesplitst met als resultaat 28.331 pagina's. Op het moment van onderzoeken waren dit alle Wob-dossiers die te vinden waren op deze webpagina. De pdf documenten zijn gedownload met behulp van *wget*, een command line tool om documenten van webpagina's te extraheren, en de Python module *BeautifulSoup*, om de URL's van de dossiers te extraheren van een HTML pagina. De pagina's (in pdf vorm) zijn op twee manieren geanalyseerd op leesbaarheid: via *pdftotext* (een *Linux* command line tool) en *Optical Character Recognition* (OCR). *Pdftotext* zet alle gevonden tekst op een pdf pagina om naar een textdocument, dat vervolgens geanalyseerd kon worden met Python om te concluderen of er leesbare tekst op een pagina werd gevonden. Bij OCR wordt het pdf document eerst omgezet naar een afbeelding om vervolgens door middel van patroonherkenning tekst te identificeren op die afbeelding.

Als tweede set van documenten zijn inventarislijsten gebruikt. De inventarislijsten zijn gezocht door eerst wederom met *BeautifulSoup* de HTML pagina's van rijksoverheid.nl/documenten af te zoeken naar URL's van Wob-dossiers. Vervolgens werden de HTML pagina's (door middel van de net gevonden URL's) van de dossiers geanalyseerd om bestandsnamen van bestanden op deze pagina te vinden. Indien de bestandsnaam blijk gaf van een inventarislijst doordat het woord 'inventaris' in de naam stond, werd het betreffende document gedownload. Op deze manier zijn er 436 inventarislijsten geïdentificeerd in 2703 dossiers. Tabellen in deze pdf documenten zijn vervolgens geëxtraheerd door middel van de Python module *pdfplumber*.

Tabel 3.1.1

Overzicht van het proces van dataverzameling

	Set 1	Set 2
Bron	wobcovid19.rijksoverheid.nl	rijksoverheid.nl/documenten
Aantal overwogen dossiers	119	2703
Aantal bestanden	28.331 losse pagina's	436 inventarislijsten
Gebruikte download-tools	BeautifulSoup, wget (Python)	BeautifulSoup, wget (Python)
Doel	Textextractie	Tabelextractie
Gebruikte extractie-tools	pdftotext, OCR	pdfplumber
Toetsing	Leesbaarheid, toegankelijkheid	Uniformiteit, compleetheid

Software-creatie

In stap 7 van de FAIRificatie is er software ontwikkeld om dossiers FAIR te kunnen produceren. Hiervoor zijn een aantal *good practices* gebruikt. Belangrijk is dat de software *open source* diende te zijn, gratis in te zien en te gebruiken voor iedereen. Hierop aansluitend is de code gedocumenteerd met behulp van *comments* om deze leesbaar te maken. Voor de interface is gebruik gemaakt van het KISS-principe (*Keep it simple, silly!*). Deze is daarom zo simpel mogelijk gehouden zodat het voor de gebruiker duidelijk is wat er kan worden verwacht.

De tool is gemaakt met een combinatie van HTML/CSS, Flask (Python) en Javascript. De interface is gemaakt in HTML en CSS. Interactieve elementen als het toevoegen van extra documenten of het downloaden van bestanden zijn geschreven in Javascript. De *back-end* van de tool is gemaakt in Python, zoals het valideren van de gegevens die zijn ingevoerd door de gebruiker. Door middel van het *Flask-framework* zijn de verschillende componenten aan elkaar verbonden.

4 Pre-FAIRificatie

Voordat er nagedacht kan worden over het toepassen van de FAIR Data Principles op Woo-dossiers is het noodzakelijk om inzicht te verkrijgen in de huidige dossiers. Aangezien de Woo ten tijde van schrijven nog maar kort van kracht is, zullen er enkel Wob-dossiers worden geanalyseerd. In stap 1 van de FAIRificatie staat de analyse van deze dossiers centraal. Aan het einde van deze stap zal duidelijk worden in hoeverre de recente Wob-dossiers al voldoen aan de FAIR Data Principles. In stap 2 kunnen er vervolgens doelen worden gesteld voor de FAIRificatie om de vindbaarheid, toegankelijkheid, uitwisselbaarheid en herbruikbaarheid van de dossiers uiteindelijk te verbeteren in de FAIRificatie. De net beschreven stappen vormen samen de pre-FAIRificatie.

4.1 Stap 1: Analyseren van de data

Stap één in het FAIRificatie proces is het verkennen en analyseren van de data die te vinden zijn in de dossiers. Jacobsen et al. (2020) stellen voor om in deze fase ‘*driving user questions*’ te gebruiken. In dit onderzoek is dat vertaald naar informatiebehoeften vanuit het perspectief van de gebruikers van de dossiers. Deze informatiebehoeften zijn geformuleerd op basis van relevante, hedendaagse voorbeelden en zijn gebruikt om problemen te identificeren met de huidige dossiers. Dit is gedaan door telkens te evalueren over de mate waarin de huidige dossiers voldoen aan de informatiebehoeften. Uiteindelijk kan er zo een conclusie worden getrokken over in hoeverre de huidige dossiers voldoen aan de FAIR Data Principles.

Probleem 1: documenten worden gepubliceerd in een slecht leesbaar en verwerkbaar formaat

Stel, een journalist, burger of andere gegadigde heeft interesse in alle verzoeken die niet binnen de wettelijke termijn van maximaal 42 dagen zijn afgehandeld. Deze informatiebehoefte stond evenals centraal in het in 2022 gepubliceerde rapport en oordeel ‘Ondraaglijk traag’ (Enthoven et al., 2022).

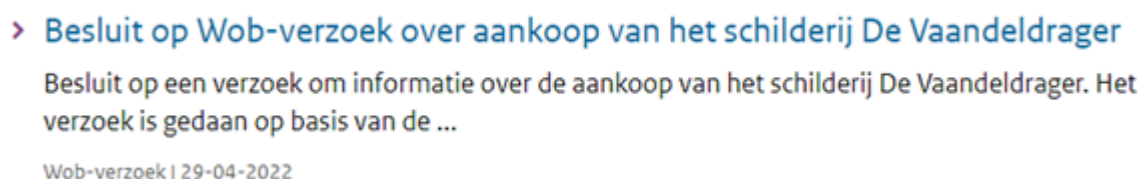
De informatiebehoefte wijst op twee benodigde stukken informatie: een verzoekdatum en een besluitdatum. Figuur 4.1.1 laat een voorbeeldpublicatie zien, gevonden op rijksoverheid.nl/documenten. De publicatie lijkt informatie over een datum te hebben meegekregen. Het is echter niet direct duidelijk om wat voor soort datum dit gaat. Vermoedelijk betreft deze datum een publicatiedatum, die niet altijd samen hoeft te vallen met de gezochte besluitdatum. Voor informatie over de specifieke datums uit de informatiebehoefte moet er in de gepubliceerde documenten worden gekeken. Figuur 4.1.2 bevat een fragment van het besluitdocument behorende bij hetzelfde verzoek. Hoewel in dit stuk tekst de benodigde datums te vinden zijn, bevat het wel acht verschillende datums. Voor een computer, dus zonder tussenkomst van een mens, is het praktisch onmogelijk om op een betrouwbare manier de juiste datums te extraheren. Zelfs met technieken als *Named Entity Recognition* (NER), aangezien er niet altijd duidelijke en uniforme hints te vinden zijn naar het type datum.

Daarnaast zijn ook niet alle documenten leesbaar en dus analyseerbaar. 28.331 pagina's uit 119 Wob-dossiers afkomstig van wobcovid19.rijksoverheid.nl zijn getoetst op computer-leesbaarheid. Door middel van de *command-line tool* 'pdftotext' werden er karakters gezocht op alle pagina's. Op 23% (6.586) van de pagina's werden geen karakters gelezen (in andere woorden, doorzoeken met control-f was niet mogelijk). Dat terwijl *Optical Character Recognition* (OCR) op vrijwel alle pagina's (99.6%) wél tekst wist te identificeren en in totaal meer dan een miljoen extra Nederlandse woorden opleverde.

Concluderend kan worden gezegd dat de documenten, die bijna altijd in een pdf formaat worden gepubliceerd, niet goed leesbaar zijn zonder tussenkomst van geavanceerde technieken als OCR. Daarnaast missen de onderzochte dossiers metadata in een verwerkbaar formaat om benodigde stukken informatie, zoals datums, gemakkelijk te kunnen vergaren. Om met de huidige dossiers toch valide resultaten in een onderzoek te verkrijgen, zullen alle besluitdocumenten met de hand moeten worden gecontroleerd, net als in het onderzoek van OSF.

Figuur 4.1.1

Voorbeeld van de publicatie van een besluit op rijksoverheid.nl/documenten.



Figuur 4.1.2

Voorbeeld van een deel van een gepubliceerd besluitdocument met gearceerde datums.

Datum 29 april 2022
Betreft Besluit Wob-verzoek inz. de Vaandeldrager

Geachte [REDACTED]

In uw brief van 5 januari 2022, ontvangen op 6 januari 2022, heeft u bij mijn ministerie met een beroep op de Wet openbaarheid van bestuur (hierna: Wob) om informatie verzocht over het schilderij De Vaandeldrager van Rembrandt van Rijn. U vraagt meer specifiek om alle interne en externe documenten als nota's, agenda's en notulen, en alle interne en externe correspondentie zoals e-mails, appberichten en sms'jes die betrekking hebben op De Vaandeldrager in brede zin, voor het tijdvak 2016 tot en met heden.

De ontvangst van uw verzoek is schriftelijk bevestigd bij brief van 12 januari 2022, kenmerknummer [REDACTED]. Bij brief van 26 januari 2022, kenmerknummer [REDACTED], is de beslistermijn met 4 weken verdaagd tot 3 maart 2022.

Op 1 maart 2022 is er telefonisch contact geweest met u over de afhandeling van uw verzoek. Daarbij heb ik aangegeven niet te kunnen voldoen aan de termijn van 3 maart 2022, in verband met de grote omvang van het aantal documenten. Er is in overleg met u toegezegd dat ik uiterlijk eind april een besluit zou nemen op uw Wob-verzoek.

Tabel 4.1.1*Kwaliteit van de inventarislijsten behorende bij Wob dossiers (N=436).*

Omschrijving	Percentage
Dossiers waar de inventarislijst een leesbare tabel bevatte *	79,4% (346)
Inventarislijst toonde bewijs van een ‘titel’ kolom **	56.2% (245)
Inventarislijst toonde bewijs van een ‘type document’ kolom **	16.3% (71)

* Tabellen zijn door middel van de Python module ‘pdfplumber’ geëxtraheerd.

** Voor het identificeren van kolomnamen is de eerste rij uit de tabel genomen waarin meer dan de helft van de kolommen gevuld is. Dit vermijdt lege rijen of kolommen met enkel een titel.

Probleem 2: inventarislijsten zijn niet altijd vindbaar en leesbaar en zijn ambigu

Gepubliceerde documenten bevatten vaak e-mail- of WhatsAppverkeer. Zo ook na een Wob-verzoek om informatie over de ‘mondkapjesdeal’ (Volkskrant, 2022). Om een reconstructie te maken van alle WhatsApp gesprekken aangaande de mondkapjesdeal kan de informatiebehoefte ontstaan naar alle documenten met WhatsApp-gesprekken en ‘mondkapjesdeal’ in de titel.

De informatiebehoefte suggereert dat er informatie moet zijn over het type document om te achterhalen welke documenten WhatsApp-gesprekken bevatten. Daarnaast moet er iets doorzoekbaars zijn als een titel om enkel documenten die gaan over de mondkapjesdeal te vinden. Tabel 4.1.1 bevat de resultaten van een onderzoek op basis van 2703 Wob-dossiers. In 436 van de dossiers werd een document gelabeld als ‘inventarislijst’ doordat het woord ‘inventaris’ in de bestandsnaam stond. Uiteindelijk voldoet 16.3% van de 436 gevonden inventarislijsten aan de gestelde eisen om de informatiebehoefte te kunnen vervullen. Met andere woorden, in 71 van de 436 inventarislijsten is er een bewijs gevonden van een kolom die het ‘type document’ aan zou geven.

Naast de matige kwaliteit van de vindbaarheid en leesbaarheid kan er nog een conclusie worden getrokken. Het geven van namen aan de kolommen in de inventarislijsten is ambigu. Tabel 4.1.2 laat voor een aantal gevonden kolomnamen de verschillende gebruikte varianten zien waarmee dezelfde stukken informatie worden bedoeld en het totaal aantal voorkomen. De uitschieter is de kolomnaam voor weigeringsgronden, daar werden 13 verschillende manieren van schrijven gevonden.

Grootschalige analyse van de vrijgegeven documenten is dus niet toegankelijk. De inventarislijsten zijn slecht te herkennen voor een computer en de informatie in wél vindbare inventarislijsten is niet goed leesbaar en er missen vaak belangrijke stukken informatie. Daarnaast worden stukken informatie op dubbelzinnige wijze geregistreerd, wat de analyse vrijwel onmogelijk maakt, tenzij dit van te voren is uitgezocht.

Tabel 4.1.2*Ambigüiteit van de gevonden kolomnamen (N=346).*

Generalisatie	% aanwezig	Gebruikte varianten
afzender	53.5% (185)	afzender; afzenders; van
beoordeling	67.3% (233)	beoordeling; beroordeling; oordeel; beoordeling-wob
datum	51.4% (178)	datum; datumdocument
documentnummer	75.1% (260)	nr; nummer; volgnummer; docnr; documentnr; id; documentnummer
ontvanger	52.0% (180)	ontvanger; ontvangers; naar; aan
titel	70.8% (245)	document; documentnaam; titeldocument; titel-doc; onderwerp; naamdocument; titel; naam
type document	20.5% (71)	soort; soortdocument; type; categorie; documenttype; typedocument; soortstuk; document-soort
weigeringsgronden	68.2% (236)	weigeringsgrond; artikelwob; wob; beslissingconform; wobgrond; uitzonderingsgrond; artikel; wobartikel; weigeringsgrondwob; weigeringsgronden; lakgrond; relevantewobgronden; grond

Probleem 3: er is geen uniforme wijze te herkennen in het registreren van informatie

Op basis van de bevindingen in het tweede probleem kan er ook gekozen worden voor een informatiebehoefte naar weigeringsgronden, die wat vaker worden gevonden in de inventarislijsten. De informatiebehoefte zou in dit geval kunnen worden geformuleerd als het vinden van alle documenten die (deels) geweigerd zijn volgens artikel 10.2.e van de Wob, ofwel de eerbiediging van de persoonlijke levenssfeer.

Bij het inspecteren van dezelfde 436 geïdentificeerde inventarislijsten als in het vorige probleem werd tegen een nieuw probleem aangelopen. Er bestaan een aantal verschillende manieren om artikel 10.2.e van de Wob aan te duiden. Met de hand zijn een aantal inventarislijsten geopend, aangezien de tabellen te verschillend zijn in vorm om automatisch te kunnen lezen. In ieder geval de volgende schrijfwijzen werden gevonden:

- 10.2.e
- Artikel 10.2.e
- 10.2.e Wob

- Artikel 10.2.e Wob
- Artikel 10 lid 2 onder e
- Artikel 10, tweede lid, aanhef en onder e Wob
- PG (afkorting persoonsgegevens)

Het blijkt dat zonder een eenduidige wijze van registreren van informatie over de weigeringsgronden het gecompliceerd is om aan de informatiebehoefte te voldoen.

De drie gevonden problemen geven aan dat de dossiers en bijbehorende documenten niet voldoen aan de FAIR Data Principes. Data zijn 1) niet vindbaar door missende informatie of niet identificeerbare of goed verstopte documenten, 2) niet toegankelijk door het niet doorzoekbare pdf formaat, 3) niet uitwisselbaar door de ambigue manier van labelen en registreren van waardes en 4) niet herbruikbaar door het gebrek aan vindbare en toegankelijke metadata. Zonder compleetheid, consistentie en uniformiteit kan er worden gesproken van *"garbage in, garbage out"*: slechte data leiden ook tot nietszeggende output.

4.2 Stap 2: Opstellen van een FAIRificatie doel

In stap 1 is geconcludeerd dat de huidige publicaties van dossiers niet aan de FAIR Data Principes voldoen. Dat betekent dat het FAIRificatie proces ervoor zal moeten zorgen dat de vindbaarheid, toegankelijkheid, uitwisselbaarheid en de herbruikbaarheid allemaal verbeteren. Hiervoor kunnen de volgende doelen worden opgesteld:

- **Doel 1:** het vinden van een kennisrepresentatie die mens- en computer-leesbaar is (vindbaarheid, toegankelijkheid)
- **Doel 2:** het identificeren van relevante metadata voor de dossiers (vindbaarheid, herbruikbaarheid)
- **Doel 3:** het ontdekken van een automatische en uniforme wijze van labelen en publiceren (uitwisselbaarheid, herbruikbaarheid)

Doel 1 focust zich vooral op het verbeteren van de vindbaarheid en toegankelijkheid door het vinden van een voor mensen en computers leesbare oplossing om de dossiers te representeren. Nadat de analyse in stap 1 heeft laten zien dat de huidige documenten in pdf formaat slecht leesbaar en toegankelijk zijn zal in stap 3, de volgende stap van de FAIRificatie aandacht worden besteed aan het definiëren van een semantisch datamodel.

Doel 2 richt zich op het identificeren van relevante metadata om de vindbaarheid en herbruikbaarheid van de dossiers te verhogen. Vanuit de analyse kan de conclusie worden getrokken dat de metadata, wanneer deze al zijn toegevoegd, vaak 'verstopt' zijn in niet uniforme inventarislijsten of slecht leesbare besluitdocumenten. In stap 4 en 5 worden metadata geïdentificeerd vanuit verschillende perspectieven en worden er regels

en beperkingen toegevoegd voor meer uniformiteit. In stap 6 worden deze FAIRificatiestappen geëvalueerd.

Doel 3 heeft als functie om de uitwisselbaarheid en herbruikbaarheid te garanderen door het zoeken naar een automatische wijze van produceren van de dossiers en een uniforme manier van publicatie van de geproduceerde dossiers. Dit zal terugkomen in de laatste stappen van het FAIRificatie-proces: stap 7 en 8.

5 FAIRificatie

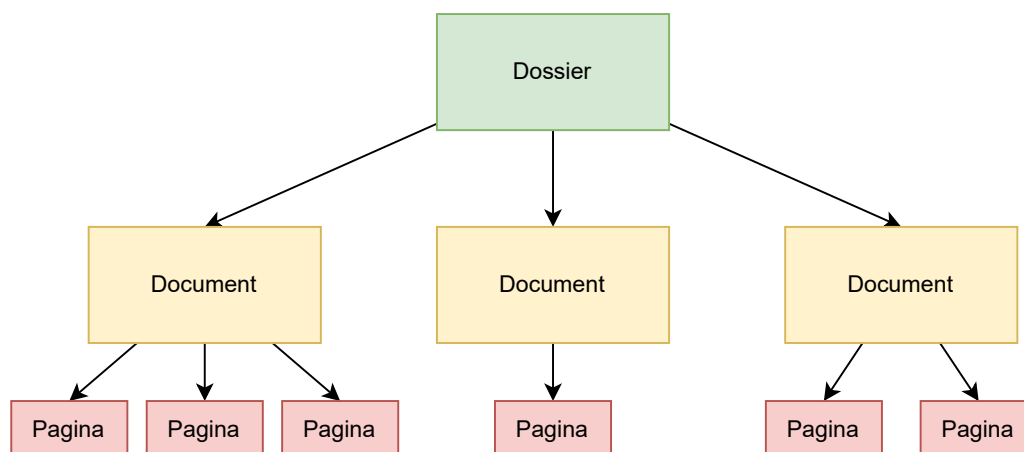
In de pre-FAIRificatie zijn de problemen met de dossiers geïdentificeerd en zijn er doelen opgesteld voor de FAIRificatie. De FAIRificatie fase bestaat uit het definiëren van een semantisch datamodel, het identificeren van relevante metadata en het opleggen van regels en beperkingen op het datamodel en de metadata.

5.1 Stap 3: Definiëren van een semantisch datamodel

De FAIRificatie begint met het definiëren van een kennisrepresentatie voor de dossiers en bijbehorende documenten. Simpler gezegd betekent dit het vinden van een goede manier om dossiers en documenten digitaal te representeren. Hiervoor worden kenmerken van gepubliceerde dossiers behandeld. Bij het analyseren van de reeds gepubliceerde Wob- of Woo-dossiers vallen een aantal zaken op. Ten eerste bestaan er verschillende soorten dossiers. Dossiers kunnen op verzoek zijn gepubliceerd of volgens de Woo actief zijn openbaar gemaakt. De dossiers kunnen niet altijd dezelfde informatie bevatten. Een actief openbaar gemaakt dossier heeft bijvoorbeeld geen verzoekdatum, terwijl een dossier op verzoek dit wel heeft. Daarnaast zijn er ook verschillende typen documenten, zoals besluiten, inventarislijsten en vrijgegeven documenten. Ook deze kunnen heel verschillende informatie bevatten. Als laatste varieert het aantal documenten per dossier flink. De net beschreven eigenschappen van de dossiers maken het lastig om de dossiers en documenten op een gestructureerde manier in tabelvorm op te slaan. De kardinaliteit, 1-op-veel, van de relatie tussen dossiers en documenten maakt het onmogelijk om de dossiers én documenten in één tabel op te slaan. De enige oplossing is het opslaan in meerdere tabellen, maar dit zou zorgen voor meerdere documenten terwijl het een stuk leesbaarder is voor mensen om één document te hebben met de gegevens om te voorkomen dat een mens de tabellen handmatig aan elkaar moet koppelen.

Figuur 5.1.1

Schematische weergave van een dossier met documenten in een boomstructuur.



De situatie vereist een wat vrijere manier van het representeren van dossiers en documenten dan een gestructureerde tabel, maar het compleet ongestructureerd opslaan is ook niet bevorderlijk voor de leesbaarheid. Er is daarom gekozen voor een semigestructureerde aanpak. Semigestructureerde data zijn gerepresenteerd als objecten in een boomachtige structuur (Buneman, 1997). Figuur 5.1.1 laat zien hoe dat er voor de dossiers en documenten ongeveer uit ziet. Dossiers bevatten een ongelimiteerd aantal documenten. De documenten bestaan op hun beurt weer uit pagina's. Javascript Object Notation (JSON) en Extensible Markup Language (XML) zijn de bekendste gestandaardiseerde gegevensformaten die een dergelijke objectrepresentatie ondersteunen. Hoewel het doel van de FAIRificatie niet is om een keuze te maken in het gegevensformaat, wordt er in latere voorbeelden gebruik gemaakt van JSON wegens de betere compatibiliteit met moderne programmeertalen als Python en Javascript (Crockford, 2006) en het open source zoekmachine softwarepakket Elasticsearch. Een versimpelde weergave van de dossiers en documenten gerepresenteerd als objecten in JSON is te zien in Codevoorbeeld 5.1.1. JSON biedt direct een manier om attributen als een titel en een datum te koppelen aan de objecten. Het grote voordeel van XML zijn de breed beschikbare schemata, voor JSON is dit een werk in uitvoering (Droettboom, 2022). Droettboom en zijn JSON Schema komen uitgebreider terug in stap 5 van de FAIRificatie.

Codevoorbeeld 5.1.1

Objectweergave van een hypothetisch dossier met documenten in JSON.

```
{
  "titel": "Wob-dossier over de verbreding van de A1",
  "datum": "2022-01-19",
  "documenten": [
    {
      "titel": "Besluitdocument",
      "type": "Besluit"
    },
    {
      "titel": "Email aangaande verbreding A1",
      "type": "Vrijgegeven document"
    }
  ]
}
```

5.2 Stap 4: Definiëren van relevante metadata

Volgens de FAIR principes 'vindbaarheid' en 'toegankelijkheid' moeten de data rijkelijk worden beschreven met relevante attributen. De metadata kunnen worden opgeslagen in het in de vorige stap gedefinieerde datamodel. Het identificeren van de relevante metadata is gedaan op vier manieren vanuit verschillende perspectieven, beginnend bij de meest generieke en eindigend met de meest Woo-specifieke attributen.

Technisch perspectief (Dublin Core)

De Dublin Core Metadata Initiative (DCMI) heeft een set met basiselementen gepubliceerd om digitale zaken te voorzien van metadata, zoals video's, afbeeldingen en documenten. Deze set is formeel gestandaardiseerd als ISO 15836 (International Organization for Standardization, 2019). De volgende attributen (9 van de in totaal 15) uit de Dublin Core kunnen worden gebruikt voor dossiers en documenten:

Dossier: identifier, title, subject, description, date, type

Document: identifier, title, description, format, date, type, language, rights

Journalistiek perspectief (Follow the Money)

Het Nederlandse onderzoeksplatform 'Follow the Money' (FTM) heeft voor journalistieke doeleinden een eigen datamodel gemaakt om 'dingen' te kunnen onderzoeken (Follow the Money, 2021). Aangezien journalisten vaak gebruik maken van de mogelijkheid tot het indienen van informatieverzoeken is het ook nuttig om vanuit dit gebruikersperspectief te redeneren. Een 'ding' is een object uit het echte leven. Onder 'dingen' vallen 'documenten'. De dossiers kunnen worden gezien als 'dingen' met daaronder vallend documenten. Uit het datamodel worden de volgende nieuwe attributen gebruikt:

Dossier: sourceUrl, retrievedAt

Document: sourceUrl, fileSize, fileName, bodyText, mimeType, fileExtension

Behandelaarsperspectief (Open State, VNG, North-Holland)

De Open State Foundation (OSF) heeft in samenwerking met de provincie Noord-Holland en de Vereniging van Nederlandse Gemeenten (VNG) een conceptversie van een handreiking genaamd 'OpenWob' gepubliceerd (Open State Foundation et al., 2021). Hierin komen ook enkele aanbevelingen voor attributen terug. De handreiking is gemaakt in samenwerking met behandelaars van de dossiers, resulterend in wat Woo-specifiekere attributen. Dit is terug te zien in de splitsing van een datum in een verzoekdatum en een besluitdatum en attributen die helpen bij het behandelen van een verzoek, zoals de verzoeker en de eerste ontvanger van het verzoek. De handreiking geeft enkel aanbevelingen op het dossierniveau.

Dossier: topicId, fileDate, decisionDate, handledBy, requester, internalId, adjourned, firstRecipient

Gebruikersperspectief

De webpagina's en de documenten van de dossiers bevatten, zoals te zien in stap 1 van de FAIRificatie, vaak al nuttige attributen die niet makkelijk vindbaar zijn voor een computer, maar wel te vinden zijn voor gebruikers. De attributen zijn gevonden door de publicatiepagina's, besluitdocumenten en inventarislijsten te analyseren. Daarnaast zijn er attributen vanuit het onderzoekersperspectief te vinden.

Publicatiepagina's. De Nederlandse overheid verrijkt publicaties meestal al met een titel, een beschrijving en een verantwoordelijk bestuursorgaan. Documenten zijn voorzien van een titel, een formaat, het aantal pagina's, de publicatiedatum en de grootte van het bestand. *Besluitdocumenten.* Dossiers bevatten altijd een besluitdocument. Hoewel niet toegankelijk en vindbaar voor een computer bevatten deze documenten vaak een verzoekdatum en een besluitdatum. Daarnaast geeft het besluitdocument aan of het verzoek verdaagd is en geeft het de definitieve beoordeling van het verzoek.

Inventarislijsten. De inventarislijsten zijn niet altijd vindbaar maar bevatten wel vaak informatie over de vrijgegeven documenten. In dezelfde 392 leesbare inventarislijsten als gebruikt in de analyse in stap 1 van de FAIRificatie zijn vijf nieuwe attributen te vinden die nog niet eerder zijn gevonden. Dit zijn de weigeringsgronden, de beoordeling, de verzender, de ontvanger en de locatie van reeds openbare stukken.

Onderzoeksperspectief Als laatste kan er vanuit de informatiebehoefte van een onderzoeker worden geredeneerd. Dit betreffen relatief eenvoudig en automatisch te vinden attributen. Een dossier kan bijvoorbeeld worden verrijkt met het aantal documenten en documenten kunnen worden aangevuld met het aantal woorden, karakters en pagina's met leesbare tekst. Om de leesbaarheid aan te geven kan er een boolean worden toegevoegd die aangeeft of de tekst te lezen is voor een computer.

De toegevoegde attributen op basis van het gebruikersperspectief zijn:

Dossier: numberDocuments, valuation, publicationDate

Document: groundsOfRefusal, valuation, originator, recipient, alreadyPublicLocation, numberTextPages, numberWords, numberCharacters, isScan

De vijf behandelde perspectieven leveren allen nuttige en relevante metadata op. De metadata zijn zo een combinatie van meer generieke, technische attributen en specifiekere attributen op basis van gebruikers en behandelaars van de Woo-dossiers. Aangezien in de tekst enkel de nieuw geïdentificeerde attributen worden benoemd, laat Tabel 5.2.1 voor alle attributen hun herkomst zien. In theorie zou het mooi zijn als al deze attributen worden verstrekt bij het publiceren van de dossiers, maar in de praktijk is dat niet haalbaar. Om te garanderen dat er toch een minimale standaard ontstaat om de uitwisselbaarheid en herbruikbaarheid te waarborgen worden er in de volgende stap regels en beperkingen opgesteld.

Tabel 5.2.1*Geïdentificeerde attributen en hun herkomst.*

a) voor dossiers					b) voor documenten			
Attribuut	DC	FTM	OSF	Wij	Attribuut	DC	FTM	Wij
identifier	x		x		identifier	x		
title	x	x	x	x	title	x	x	x
topic	x	x	x	x	description	x	x	x
topicId			x		date	x	x	x
					fileExtension	x	x	x
description	x	x	x	x	mimeType		x	
fileDate	x*	x*	x	x	documentType	x		x
decisionDate	x*	x*	x	x	language	x	x	
type	x			x	rights	x		
sourceUrl		x			sourceUrl		x	
retrievedAt		x			fileSize		x	x
handledBy			x	x	fileName		x	x
requester			x	x	bodyText		x	
internalId			x		annexType	x		x
adjourned			x	x	valuation			x
firstRecipient			x		groundsOfRefusal			x
valuation				x	alreadyPublicLocation			x
numberDocuments				x	originator			x
publicationDate				x	recipient			x
Total	7	7	12	12	numberPages			x
					numberWords			x
					isScan			x
					numberTextPages			x
					numberCharacters			x
					Total	9	10	18

Kolomnamen: DC = technisch perspectief (Dublin Core), FTM = journalistiek perspectief (Follow the Money), OSF = behandelaaersperspectief (Open State Foundation), Wij = gebruikersperspectief en onderzoekersperspectief

* In Dublin Core en het FTM datamodel komen deze datums indirect voor als normale 'date'

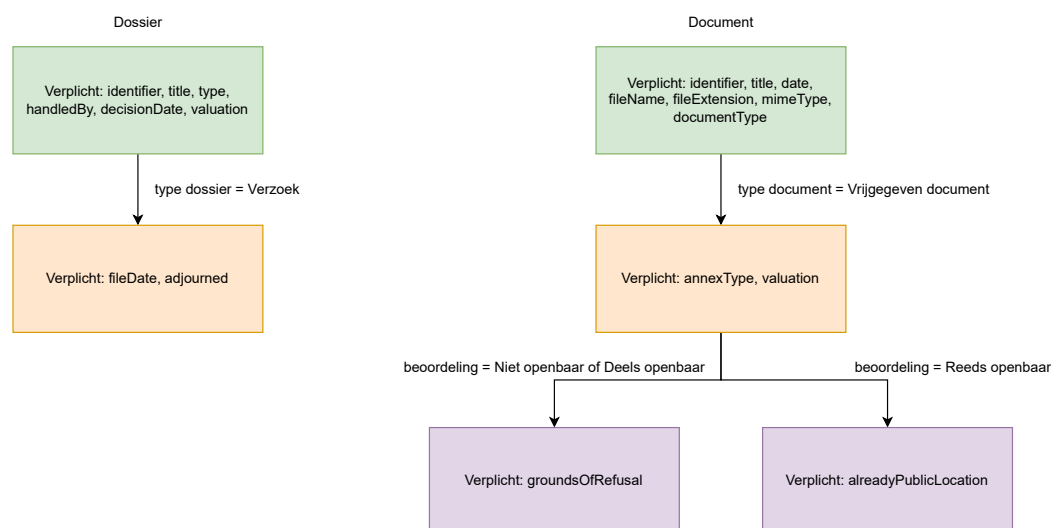
5.3 Stap 5: Opstellen van beperkingen en regels

Om de uitwisselbaarheid en herbruikbaarheid te garanderen is het van belang om beperkingen en regels te formuleren zodat alle betrokken partijen dezelfde taal spreken. In stap 1 bleek één van de problemen met de huidige dossiers het gebrek aan uniformiteit van het publiceren van metadata te zijn, zowel de attribuutnamen als de waarden. Het is daarom belangrijk om ten eerste een minimale eis te stellen aan de dossiers. Met andere woorden, het identificeren van attributen die verplicht verstrekt dienen te worden. In figuur 5.3.1 zijn de verplichte attributen te zien. Deze attributen zijn in reeds gepubliceerde dossiers op de identifier na bijna altijd te vinden. De identifier is volgens de FAIR principes een

belangrijk onderdeel om de vindbaarheid en toegankelijkheid te versterken. In het groen staan de velden die voor dossiers en documenten altijd dienen te worden verstrekt. Op basis van het type dossier of document worden er aan die lijst meer attributen toegevoegd.

Figuur 5.3.1

Verplichte attributen van dossiers en documenten.



Ook aan de waarden kunnen beperkingen en regels worden toegekend, zoals het geven van een beperkte keuze of het aanleveren van data in een specifiek formaat. In tabel 5.3.1 zijn deze beperkingen te zien. Voor alle datum gerelateerde attributen is gekozen voor het ISO 8601 formaat (International Organization for Standardization, 2019). Deze standaard beperkt de datums tot het internationaal erkende ‘YYYY-MM-DD’ formaat. ISO 639-3 wordt gebruikt om de taal van de documenten te beperken tot drie letterige land-codes, in de vorm ‘nld’ of ‘eng’ voor respectievelijk Nederlands en Engels (International Organization for Standardization, 2007). Als laatste wordt de mimeType, zoals de naam al suggereert, opgeslagen in mime type, of media type, formaat volgens IETF RFC 6838 (Internet Engineering Task Force, 2013). De beperkte keuzes zijn gebaseerd op alle mogelijke vormen die dat attribuut in het echte leven kan aannemen. Deze attributen zijn zo strakker gedefinieerd om onzinwaarden te voorkomen.

Tabel 5.3.1*Attributen van objecten en bijbehorende regel of beperking.*

Object - attribuut	Regel/Beperking
Dossier - fileDate	ISO 8601 datum formaat
Dossier - decisionDate	ISO 8601 datum formaat
Dossier - publicationDate	ISO 8601 datum formaat
Dossier - type	Beperkte keuze ('Verzoek' of 'Actieve openbaarmaking')
Dossier - valuation	Beperkte keuze ('Openbaar', 'Deels openbaar', 'Niet openbaar', 'Reeds openbaar')
Document - date	ISO 8601 datum formaat
Document - downloadDate	ISO 8601 datum formaat
Document - language	ISO 639-3 taal formaat
Document - mimeType	IETF RFC 6838 media type formaat
Document - documentType	Beperkte keuze ('Verzoek', 'Besluit', 'Inventarislijst' of 'Vrijgegeven document')
Document - valuation	Beperkte keuze ('Openbaar', 'Deels openbaar', 'Niet openbaar', 'Reeds openbaar')

Om te garanderen dat er aan de opgelegde beperkingen en regels wordt voldaan dienen de dossiers eerst gevalideerd te worden voordat deze gepubliceerd kunnen worden. In het geval van JSON kan er gevalideerd worden door een JSON Schema te creëren (Droettboom, 2022). Het schema is zelf een JSON document en bevat een definitie voor alle attributen, samen met hun opgelegde beperkingen. Het ondersteunt het aangeven van verplichte velden, het opleggen van type constraints (numberDocuments moet een nummer zijn) en het opleggen van value constraints zoals gedefinieerd in tabel 5.3.1 Met het opgestelde schema kan elk JSON document worden gevalideerd in de programmeertaal naar keuze. In Python bijvoorbeeld kan dit gedaan worden door de module 'jsonschema' te gebruiken. Codevoorbeeld 5.3.1 laat een klein fragment zien van het JSON Schema voor Woo-dossiers. De attributen zijn gedefinieerd, samen met een beperkte keuze voor het type. Ook wordt aangegeven welke attributen verplicht ingevuld moeten zijn.

Codevoorbeeld 5.3.1*Versimpeld fragment uit het JSON Schema voor Woo-dossiers.*

```

"properties": {
    "identifier": {"type": ["integer", "string"]},
    ...
    "type": {"type": "string",
             "enum": ["Verzoek", "Actieve openbaarmaking"]}
},

"required": ["identifier", "title", "handledBy", "type",
             "decisionDate", "valuation"],

```

6 Post-FAIRificatie

De laatste fase is de post-FAIRificatie. Hierin volgt in stap 6 een evaluatie. In stap 7 en 8 worden er aanbevelingen gedaan voor het automatisch produceren en uniform publiceren van de Woo-dossiers.

6.1 Stap 6: Evalueren van de FAIRificatie

In tabel 6.1.1 is de evaluatie te zien. Per FAIR principe is de situatie vóór en ná FAIRificatie geschetst om aan te geven hoe de FAIRificatie de Woo-dossiers heeft verbeterd op dat principe. Op elk principe heeft er een positieve verandering plaatsgevonden.

Tabel 6.1.1

Evaluatie van het FAIRificatie-resultaat.

FAIR Principe	Voor FAIRificatie	Na FAIRificatie
Vindbaarheid	Dossiers en documenten missen relevante metadata en/of deze zijn slecht vindbaar. Inventarislijsten zitten soms ‘verstopt’ en de pdf documenten zijn niet goed te doorzoeken.	Dossiers en documenten gebruiken een unieke identifier en zijn verrijkt met relevante attributen. De data worden opgeslagen in een bestandsformaat (JSON of XML) die de doorzoekbaarheid verhoogt.
Toegankelijkheid	Alle documenten worden in een pdf formaat verstrekt, waarvan bijna 25% niet toegankelijk en leesbaar is voor een computer. Het is lastig om automatisch bruikbare informatie te vinden in de pdf documenten.	Door de dossiers en documenten met bijbehorende metadata in een semi-gestructureerd bestandsformaat als JSON of XML op te slaan zijn ze eenvoudig te extraheren en verder te verwerken met een programmeertaal of zoekmachine-tool.
Uitwisselbaarheid	Ieder bestuursorgaan heeft een eigen wijze van labelen en publiceren. Dit resulteert in ambiguïteit voor zowel attribuutnamen als waarden.	De metadata met bijbehorende restricties zorgen voor een landelijk uniforme wijze van publiceren. Door validatie van de dossiers voldoen alle publicaties aan dezelfde standaard.
Herbruikbaarheid	De dossiers missen vindbare en toegankelijke metadata, wat de herbruikbaarheid voor onderzoek en monitoring niet ten goede komt. Ook voor de bestuursorganen zelf zorgt het gebrek aan metadata waarschijnlijk voor slechte archivering en herbruikbaarheid.	Door de identificatie van metadata vanuit het perspectief van zowel handelaars als gebruikers van de dossiers zijn de uniek te identificeren dossiers en documenten herbruikbaar voor zowel toekomstige Woo-verzoeken als journalistiek of wetenschappelijk onderzoek.

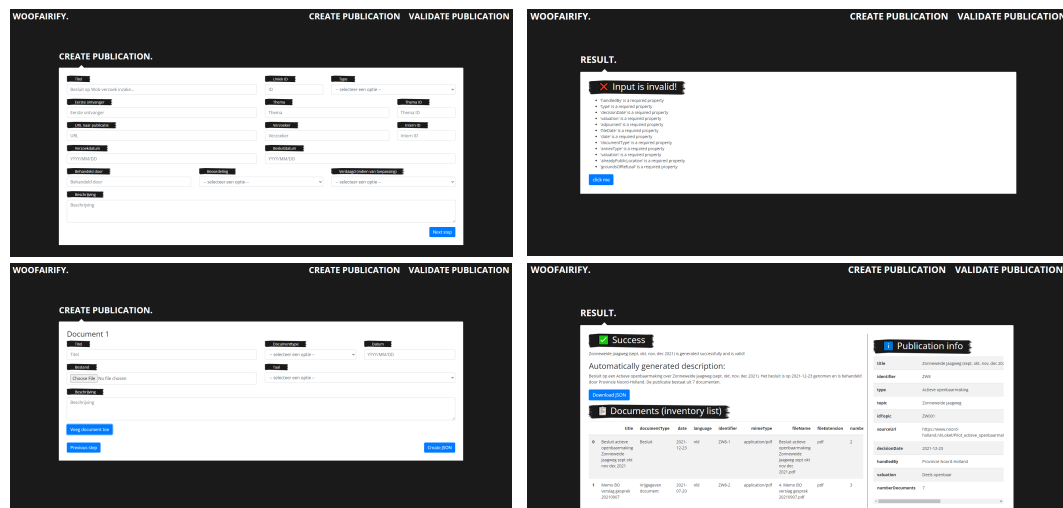
6.2 Stap 7: Het automatiseren van FAIR produceren

Met het behalen van de eerste twee doelen, het vinden van een leesbare kennisrepresentatie en het identificeren van relevante metadata, blijft het laatste doel nog over: het ontdekken van een automatische en uniforme wijze van labelen en publiceren. In deze stap zal in worden gegaan op het automatiseren van de productie van Woo-dossiers op een FAIR manier. Het produceren van dossiers is een routineklus en moet niet te moeilijk zijn en te veel tijd kosten. Voor dit doeleinde is er een tool ontwikkeld, 'WooFAIRify' (Larooij, 2022). Het doel van WooFAIRify is het bieden van een eenvoudige interface voor het invullen van relevante informatie waarna de gebruikersinput wordt gevalideerd en omgezet naar het wensbare formaat volgens de FAIRificatie, in dit geval JSON.

In Figuur 6.1.1 zijn de vier belangrijkste schermen weergegeven. Een behandelaar vult eerst in het scherm linksboven de gegevens over een dossier in, zoals een titel en een datum. Daarna kan de behandelaar documenten toevoegen (zie linksonder) met ook daarbij de metadata. Vervolgens worden de ingevulde gegevens gevalideerd. Als dat mislukt krijgt de behandelaar feedback op de ingevulde gegevens. In het geval van het scherm rechtsboven is bijvoorbeeld de behandelaar (handledBy) niet verstrekt. Als de ingevulde gegevens helemaal voldoen aan de opgestelde regels en beperkingen uit stap 5, verschijnt het resultaat in beeld. Op dit scherm kunnen een JSON van het gehele dossier en een CSV van de documenten als een inventarislijst worden gedownload. De tool maakt het mogelijk om zonder technische kennis van de attributen en het datamodel tóch op een uniforme manier dossiers te produceren.

Figuur 6.1.1

Screenshots van de ontwikkelde Woo-tool WooFAIRify.

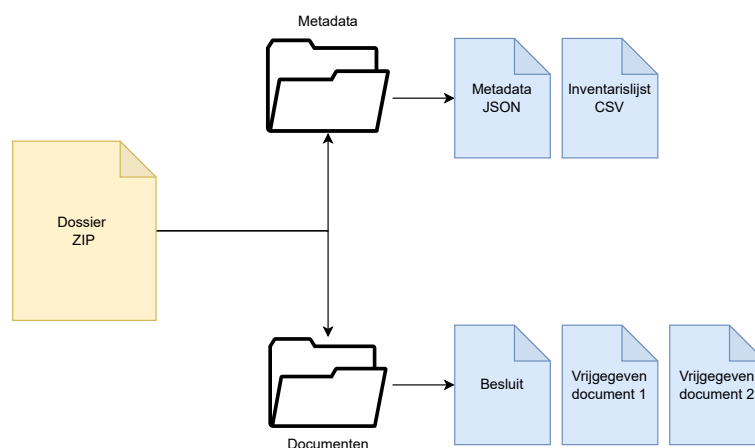


6.3 Stap 8: Identificeren van een publicatiestrategie

De laatste stap van het FAIRificatie proces geeft een handreiking voor het op een uniforme wijze publiceren van de automatisch gemaakte dossiers. Het is wenselijk om een Woo-dossier te voorzien van alle documenten die zijn vrijgegeven in pdf formaat. Daarnaast is het resultaat van de FAIRificatie een semigestructureerd document met daarin vindbare, toegankelijke, uitwisselbare en herbruikbare metadata. Eventueel kan er ook nog een inventarislijst in tabelvorm worden toegevoegd aan het dossier. Om al deze documenten te bundelen in een herkenbare vorm is het noodzakelijk om een publicatiestrategie op te stellen. Figuur 6.2.1 doet een voorzet voor een publicatiestrategie, waarin het gehele dossier als een ZIP bestand wordt verstrekt. Deze ZIP bevat vervolgens twee mappen: één map met daarin de metadata, het resultaat van de FAIRificatie, en een andere map met daarin de originele pdf documenten als reactie op een informatieverzoek of een actieve openbaarmaking.

Figuur 6.2.1

Voorbeeld van de publicatiestrategie.



Voorbeeld Zonneweide Jaagweg

Om te laten zien hoe het volledige proces van de productie van een Woo-dossier op basis van de FAIR principes eruit ziet, zal als voorbeeld een reeds gepubliceerd dossier van de provincie Noord-Holland worden gebruikt. In het licht van de vervanging van de Wob door de Woo is de provincie een pilot gestart waarbij er zoveel mogelijk informatie actief openbaar wordt gemaakt met betrekking tot de plaatsing van een zonneweide aan de Jaagweg. In de pilot wordt er gezocht naar een manier om informatie binnen één project openbaar te maken. In de huidige situatie worden de dossiers echter nog op dezelfde wijze gepubliceerd als bij de Rijksoverheid, vergelijkbaar met de bevindingen van de analyse van de dossiers in stap 1. Een FAIRificatie zal worden uitgevoerd op een dossier uit het project met daarin de maanden september tot en met december in 2021.

Figuur 6.3.1

Voorbeeld van de publicatiestrategie.

The image shows a web application interface for document management. On the left, there is a form for 'Document 1' with various input fields: Title (Zonneweide jaagweg (sept, okt, nov, dec 2021)), ID (ZW8), Type (Actieve openbaarmaking), Topic (Zonneweide jaagweg), ID Topic (ZW001), Decision Date (2021-12-23), Handled By (Provincie Noord-Holland), Valuation (Deels openbaar), and a list of documents. On the right, there is a list of documents, including 'Document 7' with details like Title (Memo BO verslag gesprek 20210520), Document Type (Vrijgegeven document), Date (2021-07-19), Recipient (Gedeputeerde Stigter, wethouder Van Dolder), Annex Type (Memo), Valuation (Deels openbaar), and Grounds of Refusal (Vrijplaats nodig openbaar).

Figuur 6.1.3 laat zien hoe de tool gebruikt kan worden om alle relevante en beschikbare gegevens over het dossier en bijbehorende documenten in te vullen. Het huidige dossier mist unieke identifiers, dus deze zijn toegevoegd als dummy-waarden. Het thema (topic) kan worden ingevuld om aan het doel om informatie binnen één project openbaar te maken te voldoen. Voor de documenten kunnen er bestanden worden toegevoegd, waarna de tool automatisch velden als de extensie, het aantal woorden, karakters en pagina's extraheert. Een fragment van het gemaakte JSON document is te zien in Codevoorbeeld 6.3.1.

Codevoorbeeld 6.3.1

Fragment van het gecreëerde JSON document van de Zonneweide Jaagweg.

```
{
  "title": "Zonneweide Jaagweg (sept, okt, nov, dec 2021)",
  "identifier": "ZW8",
  "type": "Actieve openbaarmaking",
  "topic": "Zonneweide Jaagweg",
  "idTopic": "ZW001",
  "decisionDate": "2021-12-23",
  "handledBy": "Provincie Noord-Holland",
  "valuation": "Deels openbaar",
  "documents": [
    {
      "title": "Memo BO verslag gesprek 20210520",
      "documentType": "Vrijgegeven document",
      "date": "2021-07-19",
      "language": "nld",
      "recipient": "Gedeputeerde Stigter, wethouder Van Dolder",
      "annexType": "Memo",
      "valuation": "Deels openbaar",
      "groundsOfRefusal": [
```

```

        "10.2.e"
    ],
    "identifier": "ZW8-7",
    "mimeType": "application/pdf",
    "fileName": "3. Memo BO verslag gesprek 20210520.pdf",
    "fileExtension": "pdf",
    "numberPages": 2,
    "numberCharacters": 3031,
    "numberWords": 449,
    "numberTextPages": 2
  },
  ...
]
}

```

Met de door de tool gemaakte JSON en CSV documenten kan er een voorbeeldpublicatie van het dossier worden gemaakt. In Figuur 6.3.2 is deze te zien. In twee simpele stappen - het invullen van de benodigde informatie in de tool en het creëren van een zip bestand met een overzichtelijke mappenstructuur - is er een volledig dossier gecreëerd, klaar voor publicatie.

Figuur 6.3.2

Boomstructuur van de mappen behorend bij het Zonneweide Jaagweg dossier.

```

ZW8.zip
├── Metadata
│   ├── ZW8-data.json
│   └── ZW8-inventaris.csv
└── Documenten
    ├── Besluit actieve openbaarmaking Zonneweide Jaagweg sept okt nov dec 2021.pdf
    ├── Inventarisatielijst sept okt nov dec 2021 dossier zonneweide Jaagweg.pdf
    ├── Memo B0 verslag gesprek 20210324.pdf
    ├── Memo B0 verslag gesprek 20210421.pdf
    ├── Memo B0 verslag gesprek 20210520.pdf
    ├── Memo B0 verslag gesprek 20210907.pdf
    └── Agenda B0 18112021 gedeputeerde Stigter en Koggenland zonneweide Jaagweg.pdf

```

7 Conclusie en discussie

Met de ingang van de Wet open overheid (Woo) in mei 2022 is het tijd geworden voor een kantelpunt in de wijze van verstrekken van de informatie-dossiers. Dit onderzoek is een reactie op de roep om actie tot een betere informatiehuishouding van de Nederlandse overheid. Het doel van het onderzoek is het toepassen van de FAIR Data Principes op Woo-dossiers om dossiers vindbaar, toegankelijk, uitwisselbaar en herbruikbaar te publiceren. Daarbij is er gezocht naar een antwoord op de volgende vragen:

- 1: Hoe staat het met de *FAIRness* van de door de Nederlandse overheid gepubliceerde dossiers?
- 2: Hoe zien FAIR gepubliceerde Woo-dossiers eruit?
- 3: Hoe kan software ondersteuning bieden bij het automatisch FAIR produceren en publiceren van Woo-dossiers?

Conclusies

In de pre-FAIRificatie werd bevonden dat reeds gepubliceerde dossiers niet voldoen aan de FAIR Data Principes. Dossiers bleken vaak niet vindbaar door missende informatie of 'verstopte' documenten. Door het publiceren van alle documenten, inclusief een inventaristabel, in het slecht doorzoekbare pdf formaat zijn de dossiers niet toegankelijk. Het gebrek aan uniforme en ondubbelzinnige metadata maakt het uitwisselen en hergebruiken van de dossiers praktisch onmogelijk. Uit deze punten kan geconcludeerd worden dat de FAIRness van de door de Nederlandse overheid gepubliceerde dossiers niet goed is en verandering vereist is.

Die verandering kan in de vorm van FAIR gepubliceerde Woo-dossiers. Uit de FAIRificatie blijkt een semigestructureerde manier van representeren van de dossiers de uitkomst. Dossiers en documenten zijn hierin objecten gerepresenteerd in een boomstructuur. Deze structuur is goed op te slaan in bestandsformaten als JSON of XML. Deze bestanden kunnen gevuld worden door de dossiers en documenten te voorzien van relevante attributen. Om te zorgen dat de geproduceerde Woo-dossiers uniform en dus FAIR worden gepubliceerd heeft dit onderzoek ook laten zien op welke manier er beperkingen kunnen worden opgelegd om de dossiers te kunnen valideren op compleetheid en uniformiteit. FAIR gepubliceerde Woo-dossiers zijn dus vindbaar en toegankelijk door de relevante attributen en het doorzoekbare semigestructureerde formaat. De metadata met restricties creëren een uniforme wijze van publiceren en bewaren van de dossiers. Dit maakt de dossiers ook uitwisselbaar en herbruikbaar. Een FAIR gepubliceerd Woo-dossier ziet er dan als volgt uit: een semigestructureerd document met de metadata, een inventarislijst in tabelvorm in een leesbaar formaat als CSV en daarnaast alle losse pdf documenten waarin de vrijgegeven informatie te zien is.

Software kan ondersteuning bieden bij het automatiseren van de productie van FAIR Woo-dossiers. De gecreëerde tool 'WooFAIRify' kan worden ingezet om op een gebruiksvriendelijke manier gegevens te verzamelen over de dossiers en documenten om deze vervolgens te valideren en om te zetten naar een gewenst formaat. Voor publicatie kunnen het metadata document in JSON en de inventarislijst in CSV worden gedownload

om samen met de pdf documenten te publiceren.

Discussie

Het uitgevoerde onderzoek laat zien dat de toepassing van de FAIR Data Principles op Woo-dossiers een positieve invloed heeft op de kwaliteit van de dossiers. De informatiehuishouding van de overheid gaat echter verder dan enkel Woo-dossiers. De documenten op rijksoverheid.nl gaan van kamerstukken tot jaarverslagen en beleidsnota's. Dit onderzoek heeft de weg vrij gemaakt voor verder onderzoek naar de toepassing van de FAIR Data Principles op andere overheidspublicaties. Het in dit onderzoek opgestelde FAIRificatie-proces is gericht op Woo-dossiers, maar onderzoek naar een uniform FAIRificatie-proces zou dit proces kunnen generaliseren voor gebruik op meerdere vormen van overheidsdocumenten.

Hoewel het onderzoek in theorie een positieve uitkomst laat zien, is het lastig te voorspellen hoe werkbaar de uitwerking in de praktijk zal zijn. Het bij elkaar zoeken van documenten en het formuleren van metadata blijft mensenwerk. Het werken met een standaard voor de dossiers vereist goede wil en discipline. Om te zorgen dat het verlies aan discipline niet leidt tot een mindere kwaliteit van de publicaties, wordt er in het onderzoek al wel aandacht besteed aan de validatie (stap 5), de automatische productie van Woo-dossiers (stap 7) en een wijze van uniform publiceren (stap 8). Deze aanbevelingen volgen zou de kwaliteit van de dossiers al grotendeels moeten waarborgen. Toch zal vervolgonderzoek naar de samenwerking in de praktijk moeten uitwijzen of een dergelijke standaard op nationaal niveau haalbaar is. Een ander probleem in de praktijk is de soms grote omvang van de reacties, die kunnen bestaan uit soms wel honderden documenten. Het omzetten van dergelijke reacties kan behoorlijk tijdsintensief en misschien zelfs onwenselijk zijn. Dit onderzoek richt zich op het toepassen van het FAIRificatie proces op Woo-dossiers maar doet geen praktijkonderzoek naar het werken met de standaard. Kortom, dit onderzoek levert een onderbouwde standaard, maar onderzoek vanuit behandelarsperspectief is nog nodig om eventuele problemen, zoals de omvang van dossiers, te identificeren en mogelijk te verhelpen.

Het is duidelijk dat er een verandering moet komen in de wijze van publiceren van Woo-dossiers. In dit onderzoek wordt een veelbelovende eerste stap gezet naar transparantere overheidspublicaties.

Literatuur

- Buneman, P. (1997). Semistructured data. In *Proceedings of the sixteenth acm sigact-sigmod-sigart symposium on principles of database systems* (pp. 117–121).
- Crockford, D. (2006). *Json: The fat-free alternative to xml*. Verkregen van <https://www.json.org/xml.html>
- Droettboom, M. (2022). *Understanding json schema*. Verkregen van <https://json-schema.org/understanding-json-schema/UnderstandingJSONSchema.pdf>
- Enthoven, G., Spanninga, H., Pino, C. & Spruit, A. (2021, 04). *Verbeterpunten in de informatiehuishouding voor een tijdige en kwalitatief goede afhandeling van Wob-verzoeken* (Rapport).
- Enthoven, G., Wiemers, S., den Uijl, S., Nouwen, A., Kuilman, E., Jorissen, R. & Vos-Goedhart, T. (2022, 01). *Ondraaglijk traag* (Rapport).
- Europese Commissie. (2022). *European Commission – Press release. Data Act: Commission proposes measures for a fair and innovative data economy* (Rapport).
- Follow the Money. (2021). *followthemoney documentation*. Verkregen van <https://followthemoney.readthedocs.io/en/latest/entity.html>
- GO FAIR initiative. (2022, 01). *FAIR Principles*. Verkregen van <https://www.go-fair.org/fair-principles/>
- International Organization for Standardization. (2007). *Codes for the representation of names of languages — part 3: Alpha-3 code for comprehensive coverage of languages* (ISO nr. 639-3:2007).
- International Organization for Standardization. (2019). *Date and time format* (ISO nr. 8601).
- Internet Engineering Task Force. (2013). *Codes for the representation of names of languages — part 3: Alpha-3 code for comprehensive coverage of languages* (IETF nr. RFC 6838).
- Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M. & Thompson, M. (2020). A generic workflow for the data fairification process. *Data Intelligence*, 2(1-2), 56–65.
- Larooij, M. (2022, 5). *WooFAIRify*. Verkregen van <https://github.com/maiklarooij/wooFAIRify>
- Open State Foundation, Provincie Noord-Holland & Vereniging van Nederlandse Gemeenten. (2021, 09). *Handreiking Open Wob* (Rapport).
- Reporters Without Borders. (2022). *Rsf’s 2022 world press freedom index*. Verkregen van <https://rsf.org/en/index>
- Roelfsema, K. & de Jong, K. (2020). *Onderzoek naar de stimulering van FAIR Principles bij de overheid*. (Rapport).
- Rutte, M. (2021, 01). *Kamerbrief met reactie kabinet op rapport ‘ongekend onrecht’*.
- Schultes, E. A., Jacobsen, A., Hettne, K. M., Thompson, M., Kuzak, M., Hooft, R. W., ... et al. (2019, Feb). *Essential steps of the fairification process*. OSF. Verkregen van osf.io/avrys

- van Oostveen, J. & van Loenen, B. (2014). De praktijk van de wet openbaarheid van bestuur: een gebruikersperspectief. *B&G*, 2014 (mei/juni).
- Volkskrant. (2022, 03). *Hugo de Jonge bemoeide zich actief met de mondkapjesdeal van Sywert van Lienden*. Verkregen van <https://www.volkskrant.nl/kijkverder/v/2022/hoe-hugo-de-jonge-zich-actief-bemoeide-met-de-mondkapjesdeal-van-sywert-van-lienden%7Ev497075/>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... others (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1), 1–9.
- Zuiderwijk, A. & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government information quarterly*, 31(1), 17–29.